On adaptive smoothing in kernel discriminant analysis

Anil K. Ghosh[1] **and Subhadip Bandyopadhyay**[2]

[1] **Center for Mathematics and its Applications, Mathematical Sciences Institute**

Australian National University, Canberra, ACT 0200, Australia.

[2] **Faculty of Applied Mathematics, Institute of Armament Technology,**

Girinagar, Simhagad Road, Pune 411025, India.

E-mail: anilkghosh@rediffmail.com, sdip_b_r@yahoo.com

Abstract

One popular application of kernel density estimation is in kernel discriminant analysis, where kernel estimates of population densities are plugged in the Bayes rule to develop a nonparametric classifier. Performance of these kernel density estimates and that of the corresponding classifier depend on the values of associated smoothing parameters commonly known as the bandwidths. Bandwidths that minimize mean integrated square errors of kernel density estimates often lead to poor misclassification rates in classification problems. In discriminant analysis, usually a cross validated estimate of misclassification probability is minimized to find the optimal bandwidth, and that bandwidth is used for classifying all observations. However, in addition to depending on the training data set, a good choice of bandwidth should also depend on the specific observation to be classified. Therefore, instead of fixing the value of the bandwidth parameter, in practice it may be more useful to choose it adaptively. This article presents one such adaptive classification technique, where the bandwidth is chosen based on the training sample and also on the data point to be classified. Performance of the proposed method has been illustrated using some benchmark data sets.

Keywords: Bayes' risk, bootstrap, cross-validation, kernel density estimation, MISE, misclassification rate, optimal bandwidth, scale space, spherical symmetry.

1 Introduction

Kernel density estimation (see e.g., Silverman, 1986; Scott, 1992, Wand and Jones, 1995) is a popular method for constructing nonparametric estimates of population densities. If $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are d-dimensional observations from the j^{th} $(j = 1, 2, \ldots, J)$ population, the kernel estimate of the density function $f_j(\mathbf{x})$ 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 $K(\cdot)$ is a density function on the d-dimensional space, and $h > 0$ is a smoothing parameter popularly known as the bandwidth. In kernel discriminant analysis (see e.g., Hand, 1982; Duda, Hart and Stork, 2000; Hastie, Tibshirani and Friedman, 2001), one plugs in these density estimates into the Bayes rule (see e.g., Anderson, 1984; McLachlan, 1992) to construct a classifier, which assigns an observation $\mathbf{x}$ into the class having the maximum estimated posterior probability. The resulting decision rule $\mathsf{d}_h(\mathbf{x})$ can also be expressed as

$$\mathsf{d}_h(\mathbf{x}) = \arg\max_j \pi_j \hat{f}_{jh}(\mathbf{x}),$$

where π_j is the prior probability of the j^{th} class $(j = 1, 2, \ldots, J)$. Various choices of kernel functions are available in the literature (see e.g., Silverman, 1986). Throughout this article, without mentioned otherwise, we will use the Gaussian kernel $K(\mathbf{t}) = (2\pi)^{-d/2} e^{-\mathbf{t}'\mathbf{t}/2}$ for our purpose.

It is quite transparent that the kernel density estimates $\hat{f}_{1h}(\mathbf{x}), \hat{f}_{2h}(\mathbf{x}), \ldots, \hat{f}_{Jh}(\mathbf{x})$ and the corresponding decision rule $\mathsf{d}_h(\mathbf{x})$ change with the bandwidth parameter h. So, changes in the values of the bandwidth may result in substantial change in the misclassification rate of the resulting classifier (see e.g., Hand, 1982; Scott, 1992, Ghosh and Chaudhuri, 2004). Therefore, in a classification problem, it is very important to choose the bandwidth appropriately. In usual density estimation problems, the optimal bandwidth is generally taken to be the one that minimizes the mean integrated square error (MISE$(h) = E[\int \{\hat{f}_{jh}(\mathbf{x}) - f_j(\mathbf{x})\}^2 \, d\mathbf{x}]$, see e.g., Silverman,1986;

Scott,1992) of the kernel density estimate, and it is a common practice to use the same bandwidth when such density estimates are used for classification. Usually, one uses the least square cross validation technique (see e.g., Hall, 1983; Silverman, 1986) to find a data based estimate of MISE and that is minimized over h to find out the optimal bandwidth. However bandwidths chosen in this way often lead to poor performance in classification problems (see e.g., Ghosh and Chaudhuri, 2004). Several other methods for bandwidth selection are available in the literature (see e.g., Muller, 1984; Stone, 1984; Hall *et.al.*, 1991; Sheather and Jones, 1991; Wand and Jones, 1995; Jones, Marron and Sheather 1996) but most these methods have their focus on the accuracy of the density estimates rather than minimizing the misclassification rate. But, in discriminant analysis, our prime interest is to minimize the average misclassification probability, which is given by

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

One may use cross validation techniques (see e.g., Lachenbruch and Mickey, 1968; Stone, 1977; Ripley, 1996) to estimate these misclassification rates for different h in order to find out the optimum one. But these cross-validation methods use naive empirical proportion to estimate Δ, and as a result they often lead to multiple values of h as the minimizers of the error rate, from which it is difficult to find out the optimum bandwidth. Ghosh and Chaudhuri (2004) proposed to minimize a smooth estimate of misclassification rate to find a unique optimum bandwidth for classification. Hall and Kang (2005) pointed out some limitations of cross-validation approach, and they proposed to use a bootstrap estimate of Δ for bandwidth selection. Hall and Wand (1988) suggested to estimate the optimal bandwidth in two-class classification problems by minimizing the MISE of the density difference.

However, all these bandwidth selection methods use some criterion to select the optimum bandwidth from the training data set, and then use that selected bandwidth for classification of all observations. But one should note that in addition to its dependence on the training data, a

good choice of bandwidth may also depend on the specific observation to be classified. A fixed value of bandwidth may not work well for classifying observations in all parts of the measurement space. Therefore, instead of fixing the value of bandwidth over the entire sample space, in practice it may be more useful to choose it adaptively. This article presents one such adaptive classification technique based on kernel density estimates, where instead of using any pre-specified value of the bandwidth parameter, it is selected adaptively based on the observation to be classified. Though several methods for adaptive bandwidth selection are available in the literature (see e.g., Breiman, Meisel and Purcell, 1977; Abramson, 1982; Terrel and Scott, 1992; Sain and Scott, 1996), almost all of them concentrate more on the accuracy of the density estimates rather than minimizing the misclassification rate in a classification problem. As a consequence, like the classifier that uses optimum MISE bandwidth, the resulting classifier often leads to poor performance, which we will see later in Section 4.

The organization of the paper is as follows. In Section 2, we start with an illustrative example which shows the importance of adaptive bandwidth selection in classification problems. In this section, we also carry out some theoretical analysis to investigate the behavior of the conditional misclassification probability for varying choices of bandwidth. Detail description of the data dependent adaptive bandwidth selection technique is given in Section 3. In Section 4, we use some benchmark data sets to compare the performance of the proposed method with those of the traditional ones. Finally, Section 5 contains a brief summary of the work and related discussions.

2 Motivation

Let us consider an example with two-class classification problem, where both classes are bivariate normal having the same dispersion matrix $\mathbf{I_2}$ (the identity matrix) but different location parameters (0,0) and (2,2). We consider the same sample size 50 for each class, while prior probabilities for the

two classes are taken as 0.55 and 0.45, respectively. In this example, because of spherical symmetry of the population distributions, same bandwidth can be used for all co-ordinate variables. Again, since the populations have the same distributional structure and the same number of observations, it is quite reasonable to use the same bandwidth for these two classes. If h is taken as that common bandwidth, the conditional misclassification probability at $\mathbf{x}$ can be expressed as

$$\psi_h(\mathbf{x}) = p(1 \mid \mathbf{x})\ P\{\pi_1 \hat{f}_{1h}(\mathbf{x}) < \pi_2 \hat{f}_{2h}(\mathbf{x})\} + p(2 \mid \mathbf{x})\ P\{\pi_1 \hat{f}_{1h}(\mathbf{x}) > \pi_2 \hat{f}_{2h}(\mathbf{x})\},$$

where $p(1 \mid \mathbf{x}) = \pi_1 f_1(\mathbf{x})/\{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})\}$ and $p(2 \mid \mathbf{x}) = \pi_2 f_2(\mathbf{x})/\{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})\}$ are true posterior probabilities of two classes at $\mathbf{x}$. Usually, the probability function P does not have a close form expression. One can approximate it using large scale Monte Carlo simulations. We generate 10,000 samples each of size 100 (50 from each class) to approximate $\psi_h(\mathbf{x})$ for different values of h. These estimates and their gray scale values are plotted in Figure 1 for four different choices (0.80, 0.80), (0.95, 0.95), (1.10, 1.10) and (1.25, 1.25) of $\mathbf{x}$, which are marked as A, B, C and D, respectively.

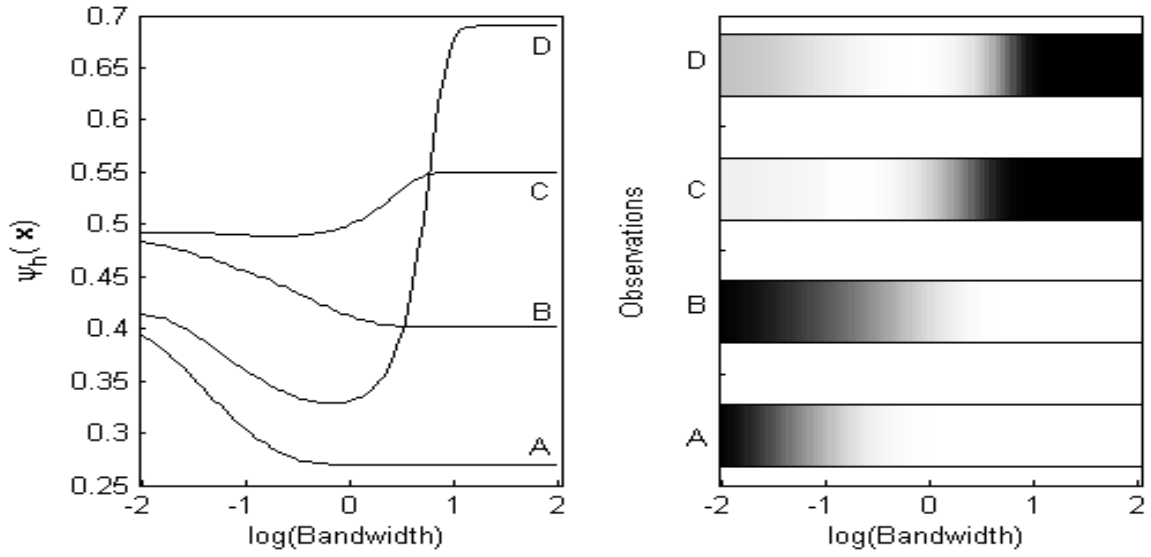

Figure 2.1 : Average misclassification rates for different bandwidths and different values of x ($\pi_1 \neq \pi_2$).

In this figure, we observe different behaviors of ψ_h at these four different points [see Figure 2.1(a)]. For $\mathbf{x} = (0.80, 0.80)$ and $(0.95, 0.95)$, $\psi_h(\mathbf{x})$ gradually decreases as h increases but for

$\mathbf{x} = (1.10, 1.10)$, the curve shows nearly an opposite behavior. For $\mathbf{x} = (1.25, 1.25)$, it decreases initially but then grows up and attains the maximum. For these four data points, ψ_h is minimized at 1.70, 2.43, 0.52 and 0.83, respectively, which can be considered as optimal bandwidths for classifying these observations. Grey scale values of ψ_h (re-scaled to have the minimum value 0 and the maximum value 1) in Figure 2.1(b) tells the same story more clearly. Here white color denotes the minimum value 0 and black color denotes the maximum value 1. Intensity of the color varies with the magnitude of re-scaled ψ_h function. Clearly, one would like to use the bandwidths in the white region for classifying an observation. Different behavior of ψ_h for different $\mathbf{x}$ is quite transparent from this figure as well. Bandwidths that are good for classifying A and B are certainly not the best choice for classification of other two data points (C and D). Ideally, one should use some large bandwidth for classification of A and B, while small values of h should be used for classifying C and D. This example, though a very simple one, clearly demonstrates the necessity of adaptive bandwidth selection in kernel discriminant analysis.

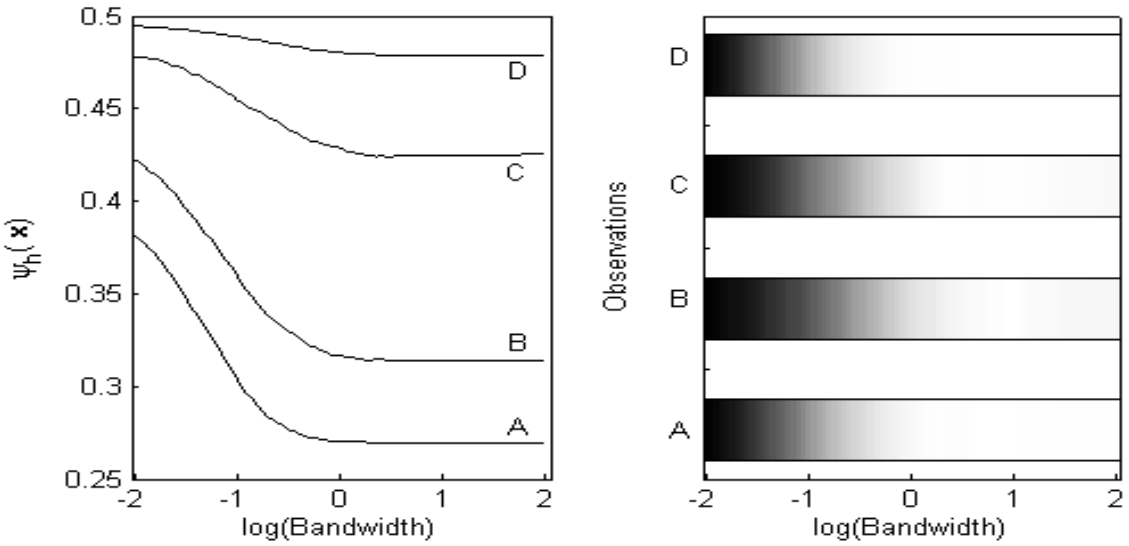

Figure 2.2 : Average misclassification rates for different bandwidths and different values of x ($\pi_1 = \pi_2$).

However, the figure gets completely changed when the prior probabilities are taken to be equal (see Figure 2.2). In that case, for all four values of $\mathbf{x}$, we observe that the misclassification

rate gradually decreases up to a certain level as the bandwidth increases, and then after reaching its minimum it becomes almost flat. As a result, any large bandwidth seems to be a good choice for classification of all these four observations. In this problem, for A, B, C and D, minimizers of ψ_h were found to be very close; 2.38, 2.13, 2.36 and 2.02, respectively. But in all these cases use of any large bandwidth led to near optimal values of ψ_h. From our knowledge on density estimation, we know that the use of large bandwidth reduces the variance of the kernel density estimates but it increases the bias and generally the MISE of the density estimates as well (see the proof of Theorem 2.1 in Appendix for expressions of expectation and variance of $\hat{f}_{jh}$ for large h). Therefore, large bandwidths are not expected to yield good result for density estimation. But in this classification problem we observe them to perform well. This counter intuitive behavior of ψ_h can be explained using Figure 2.3.

Here we generate a test set of size 1000 taking 500 observations from each class and classify them using a training set of size 100 (50 from each class). We consider the same parameters for the two populations that we used in Figure 2.1 and Figure 2.2. Figure 2.3(a) shows the true and the estimated posterior probabilities for class-1 when the optimal MISE bandwidth is used for classification. Though high concentration of points near the $x = y$ line clearly indicates very little bias for posterior probability estimates, because of higher variability, we observe few points on the second and the fourth quadrants indicating misclassifications. On the other hand, when a large bandwidth ($h = 3$) is used for classification, posterior estimates become very close to 0.5 (which is quite evident from the expression of $E\{f_{jh}(\mathbf{x})\}$ and $Var\{f_{jh}(\mathbf{x})\}$ in the Appendix) leading to more bias but this large bandwidth reduces the variance of the posterior estimates. As a result, the number of points in the second and the fourth quadrants gets reduced as well (see Figure 2.3(b)). When optimum MISE bandwidth was used, mean square error (MSE) of the posterior probability estimate was found to be 0.0140 but it misclassified 86 observations, whereas $h = 3$ led to an MSE of 0.1082, while it misclassified 81 observations. The picture becomes more clear in higher dimensions.

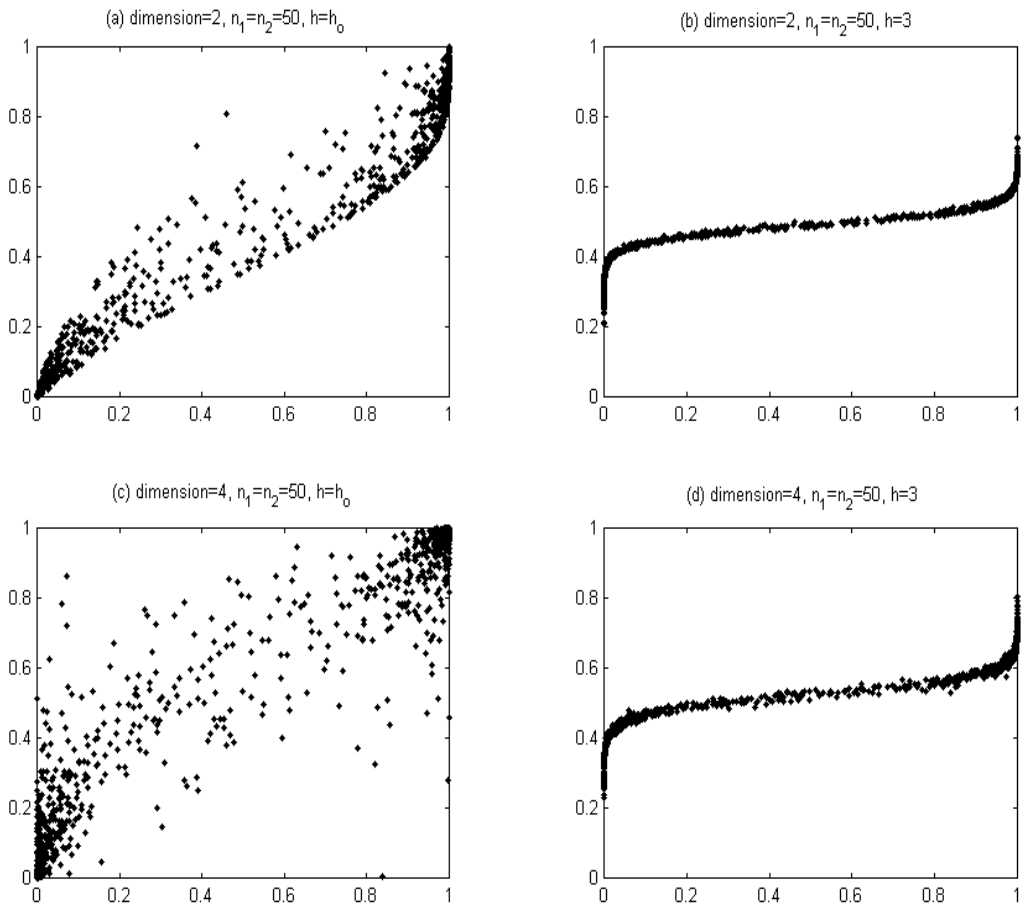

Figure 2.3 : True (x-axis) and estimated (y-axis) posterior probabilities for two different choices of bandwidths.

In Figure 2.3(c) and 2.3(d), we consider a four-dimensional problem, where two independent $N(0,1)$ noise are added in each class as two new measurement variables. In this example, optimum MISE bandwidth led to an MSE of 0.0197 but it misclassified 103 observations, while use of $h = 3$ increased the mean square error to 0.1138 but reduced the number of misclassification to 90. Friedman (1997) studied the effect of bias and variance of posterior estimates on misclassification rate, where he gave a theoretical argument to show that for classification low variance of posterior estimates may be more important than their bias. We also observed the same phenomenon here. Similar discussion on bias variance trade off in classification is also available in Ghosh, Chaudhuri and Sengupta (2006).

For varying choices of bandwidth h, following the ideas and terminology in Chaudhuri and Marron (1999, 2000), $E\{\hat{f}_h(\mathbf{x})\}$ and $\hat{f}_h(\mathbf{x})$ can be called 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 for any fixed bandwidth h, the variance of a kernel density estimate (which is an average of a set of i.i.d. random variables) shrinks to zero as the sample size increases, and as a consequence, the distribution of $\hat{f}_h(\mathbf{x})$ tends to be degenerate at $E\{\hat{f}_h(\mathbf{x})\}$.

Therefore, when the prior probabilities of the competing populations are all equal, for any fixed value of h, as the sample sizes $n_1, n_2, \ldots, n_J$ tend to infinity, the kernel density estimate based classifier tends to classify an observation to 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 (see Ghosh and Chaudhuri, 2004). So, under this set up for any fixed value of h, $\mathsf{d}_h(\mathbf{x})$ has a point wise convergence to the optimum Bayes rule, and the corresponding misclassification rate $\Delta(h)$ asymptotically converges to the optimal Bayes risk. It should be noted that for larger values of h, as because the variance of $\hat{f}_{jh}(\mathbf{x})$ converges to zero rather quickly, this convergence to Bayes risk is much faster there.

Notice that in our example, the density functions f_1 and f_2, and the kernel K were spherically symmetric, while f_1 and f_2 also satisfied a location shift model. As we have discussed in the last paragraph, this is an ideal situation for a kernel density estimate based classifier with large bandwidth to perform well when the prior probabilities are equal, and that is what we observed in Figure 2. So, in this set up bandwidth selection is less important, and instead of adopting any bandwidth selection technique one can arbitrarily choose any large bandwidth to achieve good misclassification rate. However, the scenario becomes completely different when the population

distributions do not satisfy any symmetry condition or they have different prior probabilities (see Figure 1). In the unequal prior case, large bandwidth did not seem to be a good choice for classifying all observations. In fact, unlike the equal prior case, there was no bandwidth which performed equally well for all observations. Optimum bandwidths for these four observations turned out to be very different, which shows the necessity for adaptive bandwidth selection.

In a J-class problem, if one uses the same bandwidth for all populations, for a given value of $\mathbf{x}$, the conditional misclassification probability is given by

$$\begin{aligned}\psi_h(\mathbf{x}) &= \sum_{j=1}^{J} p(j \mid \mathbf{x})\ P\{\pi_j \hat{f}_{jh} < \pi_i \hat{f}_{ih} \text{ for some } i \neq j\} \\ &= \frac{1}{G(\mathbf{x})} \sum_{j=1}^{J} \pi_j f_j(\mathbf{x})\ P\{\pi_j \hat{f}_{jh} < \pi_i \hat{f}_{ih} \text{ for some } i \neq j\},\end{aligned}$$

where $G(\mathbf{x}) = \sum \pi_j f_j(\mathbf{x})$ is the density of the mixture of J distributions at $\mathbf{x}$. Minimizing the $\psi_h(\mathbf{x})$ function w.r.t. h one gets the adaptive bandwidth which is to be used for classifying the observation $\mathbf{x}$. Note that instead of $\psi_h(\mathbf{x})$, one can also minimize the overall misclassification probability $\Delta(h) = \int \psi_h(\mathbf{x}) G(\mathbf{x}) d\mathbf{x}$ to get a single optimum bandwidth for classification (see e.g., Ghosh and Chaudhuri, 2004; Hall and Kang, 2005). But as we have demonstrated in Figure 1, a single bandwidth often fails to work well for classification of all observations. Note that $\min_h \Delta(h) \geq \int \{\min_h \psi_h(\mathbf{x})\} G(\mathbf{x}) d\mathbf{x}$ and hence adaptive choice of bandwidth parameter is expected to perform better in terms of misclassification rates.

We have already discussed about an interesting asymptotic behavior of $\psi_h(\mathbf{x})$ for large h in the equal prior case under a spherical symmetry condition. However, in practice, the population distribution may not satisfy any symmetry condition or any location shift model. The following theorem throws some light on the large sample behavior of $\psi_h(\mathbf{x})$ for large h under a more general set up when the population densities may or may not satisfy any symmetry condition.

Theorem 2.1: *Suppose that the population densities satisfy the condition* $\int \|\mathbf{x}\|^6 f_j(\mathbf{x}) d\mathbf{x} < \infty$ *for* $j = 1, 2, \ldots, J$. *Also assume that the kernel function* K *has a mode at* $\mathbf{0}$ *and it has bounded*

third derivative. Further assume that $1 \leq m \leq J$ *and* $\pi_{j_1} = \pi_{j_2} = \ldots = \pi_{j_m} > \pi_i$ *for all* $i \notin \{j_1, j_2, \ldots, j_m\}$. *Then as* $\min\{n_1, n_2, \ldots, n_J\}$ *and the bandwidth* h *both tend to infinity,* $\psi_h(\mathbf{x})$ *converges to 1-* $\pi_j f_j(\mathbf{x})/G(\mathbf{x})$ $[= 1 - p(j \mid \mathbf{x})]$, *where*

$$j = arg \max_{k \in \{j_1, j_2, \ldots, j_m\}} \{\boldsymbol{\mu}_k^{'} A \boldsymbol{\mu}_k - 2\mathbf{x}^{'} A \boldsymbol{\mu}_k + trace(A\boldsymbol{\Sigma}_k)\} \quad for\ A = \nabla^2 K(\mathbf{0}),$$

where $\boldsymbol{\mu}_k$ *and* $\boldsymbol{\Sigma}_k$ *are the mean vector and the variance-covariance matrix of the* k*-th class.*

Note that in a two class problem, being a convex combination of $p(1 \mid \mathbf{x})$ and $p(2 \mid \mathbf{x})$, for all values of h, $\psi_h(\mathbf{x})$ lies between these two posterior probabilities. For classifying $\mathbf{x}$, one aims to choose the adaptive bandwidth h for which $\psi_h(\mathbf{x})$ is minimum. Now, from the above theorem it is quite clear that when $\pi_1 > \pi_2$, whatever be the value of $\mathbf{x}$, for large h, $\psi_h(\mathbf{x})$ tends to $p(2 \mid \mathbf{x})$. So, when $p(1 \mid \mathbf{x}) > p(2 \mid \mathbf{x})$, use of large bandwidth leads to the lowest value of $\psi_h(\mathbf{x})$, which indicates that large bandwidths are good for classifying $\mathbf{x}$. On the other hand, when $p(1 \mid \mathbf{x}) < p(2 \mid \mathbf{x})$, use of large bandwidth is not a sensible option as it leads to the highest value of the conditional misclassification probability $\psi_h(\mathbf{x})$. This is what we observed in Figure 2.1. For observations A and B, where $p(1 \mid \mathbf{x})$ was larger than $p(2 \mid \mathbf{x})$, use of large bandwidth led to good performance but in the other two cases (C and D), where $p(1 \mid \mathbf{x})$ was smaller than $p(2 \mid \mathbf{x})$, the performance of the large bandwidth classifier was not satisfactory at all.

Also, it is easy to notice that when the population distributions satisfy a location shift model, $\boldsymbol{\Sigma}_j$ is same for all populations. Again, if they are spherically symmetric, $f_i(\mathbf{x}) > f_j(\mathbf{x}) \Leftrightarrow \|\mathbf{x} - \boldsymbol{\mu}_i\| < \|\mathbf{x} - \boldsymbol{\mu}_j\| \Leftrightarrow (\mathbf{x} - \boldsymbol{\mu}_i)^{'} A (\mathbf{x} - \boldsymbol{\mu}_i) < (\mathbf{x} - \boldsymbol{\mu}_j)^{'} A (\mathbf{x} - \boldsymbol{\mu}_j)$ (since $A = \nabla^2 K(\mathbf{0})$ is negative definite) $\Leftrightarrow \boldsymbol{\mu}_i^{'} A \boldsymbol{\mu}_i - 2\mathbf{x}^{'} A \boldsymbol{\mu}_i > \boldsymbol{\mu}_j^{'} A \boldsymbol{\mu}_j - 2\mathbf{x}^{'} A \boldsymbol{\mu}_j$. Therefore, when the priors are equal, for large h, $\psi_h(\mathbf{x})$ converges to $1 - p(j \mid \mathbf{x})$, where $j = \arg\max f_j(\mathbf{x}) = \arg\max p(j \mid \mathbf{x})$. So, for all $\mathbf{x}$, $\psi_h(\mathbf{x})$ asymptotically reaches its lower bound, and hence the misclassification rate of the resulting classifier converges to the optimal Bayes risk as we have discussed before.

3 Data analytic implementation

Since the expression of ψ_h involves unknown population densities $f_1, f_2, \ldots, f_J$, it is not possible to find out the adaptive bandwidth by minimizing ψ_h with respect to h. In practice, one needs to find a suitable data based estimate for ψ_h and to minimize it. Note that instead of minimizing ψ_h, one can forget about the $G(\mathbf{x})$ term (which does not involve any h) in the denominator and estimate the optimum bandwidth parameter by minimizing

$$\begin{aligned}\psi_h^*(\mathbf{x}) &= \sum_{j=1}^{J} \pi_j f_j(\mathbf{x}) P\{\pi_j \hat{f}_{jh} < \pi_i \hat{f}_{ih} \text{ for some } i \neq j\} \\ &= \sum_{j=1}^{J} \pi_j f_j(\mathbf{x}) \left[1 - P\{\pi_j \hat{f}_{jh} \geq \pi_i \hat{f}_{ih} \text{ for all } i \neq j\}\right]\end{aligned}$$

Given the value of $\mathbf{x}$, since the kernel density estimates are independently distributed, $\psi_h^*(\mathbf{x})$ can be re-written as

$$\psi_h^+(\mathbf{x}) = \sum_{j=1}^{J} \pi_j f_j(\mathbf{x}) \left[1 - \int \left(\prod_{i \neq j} P\{\pi_i \hat{f}_{ih}(\mathbf{x}) \leq u\}\right) g_{jh}(u)\, du\right],$$

where $g_{jh}(\cdot)$ is the p.d.f. of $\pi_j \hat{f}_{jh}(\mathbf{x})$. The probability function P in the above expression of ψ_h^+ does not have a closed form expression. One can use resampling techniques like bootstrap (see e.g., Efron, 1982; Efron and Tibshirani, 1993) to estimate this probability. However, bootstrap technique is computationally expensive. If not impossible, it is computationally very difficult to use this technique even when we have moderately large training samples. Note that given the values of h and $\mathbf{x}$, kernel density estimates are averages of i.i.d. random variables. Therefore, when the training data set is large or moderately large, it is quite reasonable to use normal approximation to the distribution of kernel density estimates for estimating this probability function. As a result, $\psi_h^+(\mathbf{x})$ can be approximated by

$$\psi_h^\circ(\mathbf{x}) = \sum_{j=1}^{J} \pi_j f_j(\mathbf{x}) \left[1 - \int \left(\prod_{i \neq j} \Phi\left\{\frac{u - \pi_i \mu_{ih}(\mathbf{x})}{\pi_i s_{ih}(\mathbf{x})}\right\}\right) \phi\left\{u, \pi_j \mu_{jh}(\mathbf{x}), \pi_j s_{jh}(\mathbf{x})\right\}\, du\right],$$

where $\mu_{jh}(\mathbf{x}) = \int \hat{f}_{hj}(\mathbf{x}) f_j(\mathbf{x}) d\mathbf{x}$ and $s_{jh}^2(\mathbf{x}) = \int \hat{f}_{hj}^2(\mathbf{x}) f_j(\mathbf{x}) d\mathbf{x} - \mu_{jh}^2(\mathbf{x})$ are the mean and the variance of $\hat{f}_{jh}(\mathbf{x})$ $(j = 1, 2, \ldots, J)$, $\Phi(\cdot)$ 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. The expression of $\psi_h^\circ(\mathbf{x})$ again involves unknown density functions $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_J(\mathbf{x})$, and the means $\mu_{jh}(\mathbf{x})$ and the variances $s^2_{jh}(\mathbf{x})$ of kernel density estimates depend on the unknown population densities as well. To get a reasonable estimate for ψ_h°, we plug-in the kernel density estimate (see e.g. Silverman, 1986; Scott, 1992) of $f_j(\mathbf{x})$ in the expression of $\mu_{jh}(\mathbf{x})$ and $s^2_{jh}(\mathbf{x})$, where the bandwidth that minimizes the mean integrated square error (MISE) is used for this density estimation. Least square cross validation technique (see e.g., Hall, 1983; Silverman, 1986) is used to find out this optimum MISE bandwidth. When Gaussian kernels are used, these estimated means $\hat{\mu}_{jh}(\mathbf{x})$ and variances $\hat{s}^2_{jh}(\mathbf{x})$ have nice closed form expressions given by

$$\hat{\mu}_{jh}(\mathbf{x}) = \frac{1}{n_j}\sum_{i=1}^{n_j} \phi_d\left(\mathbf{x}, \mathbf{x}_{ji}, \{h^2 + h^2_{\circ j}\}\mathbf{I}_d\right),$$
$$\hat{s}^2_{jh}(\mathbf{x}) = \frac{1}{n_j - 1}\left[\left(\frac{1}{4\pi h^2}\right)^{d/2}\left\{\frac{1}{n_j}\sum_{i=1}^{n_j} \phi_d\left(\mathbf{x}, \mathbf{x}_{ji}, \{0.5h^2 + h^2_{\circ j}\}\mathbf{I}_d\right)\right\} - \hat{\mu}^2_{jh}(\mathbf{x})\right],$$

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 j^{th} population. We also plug-in the kernel density estimate $\hat{f}_{jh_{\circ j}}$ for f_j $(j = 1, 2, \ldots, J)$ in the expression of $\psi_h^\circ(\mathbf{x})$. As a result, the data based estimate for $\psi_h^\circ(\mathbf{x})$ is obtained as

$$\hat{\psi}_h^\circ(\mathbf{x}) = \sum_{j=1}^{J} \pi_j \hat{f}_{jh_{\circ j}}(\mathbf{x})\left[1 - \int\left(\prod_{i\neq j}\Phi\left\{\frac{u - \pi_i\hat{\mu}_{ih}(\mathbf{x})}{\pi_i\hat{s}_{ih}(\mathbf{x})}\right\}\right)\phi\left\{u, \pi_j\hat{\mu}_{jh}(\mathbf{x}), \pi_j\hat{s}_{jh}(\mathbf{x})\right\}\, du\right].$$

The integral appearing in the above expression can be evaluated numerically without much difficulty and to a great degree of accuracy. Note that, $\hat{\psi}_h^\circ(\mathbf{x})$ is a smooth function, and we propose to minimize it over h to find the optimal adaptive bandwidth for classifying $\mathbf{x}$.

Use of a simplified Gaussian kernel with a single h in all directions requires some preliminary transformation of the data. In this article, for all data analytic purpose, we used the usual moment based estimate of the dispersion matrix for this transformation. This standardization makes the data cloud spherical in some sense and thereby makes the use of the same h in all directions more justified. Of course, one may use other types of standardization as well (see e.g., Cooley and

MacEachern, 1998).

4 Results on benchmark data sets

In this section, we use some benchmark data sets to evaluate the performance of the proposed method. Along with the misclassification rates of our proposed classifier, we also report the error rates for the kernel classifier that uses the bandwidths that minimize estimated MISEs of the kernel density estimates (see e.g., Silverman, 1986) and the classifier that uses the optimum bandwidth that minimizes the smooth estimate of average misclassification probability as proposed by Ghosh and Chaudhuri (2004). For future reference we will refer to these two classifiers as the MISE bandwidth classifier and the optimal bandwidth classifier, respectively. Unlike our proposed method, these two kernel classifiers use a fixed bandwidth for classifying all observations. To facilitate our comparison, we also report the misclassification rates of kernel density estimate based classifier that uses adaptive bandwidth for density estimation. We use Abramson's (see Abramson, 1982) square root law for this purpose. This square root rule is very simple and easy to implement. It is basically a two-stage procedure. At the first stage it computes a global bandwidth $h_{\circ j}$ by minimizing MISE of the kernel density estimate and then use that bandwidth to compute local bandwidths at the data points $\mathbf{x}_{ji}, i = 1, 2, \ldots, n_j$. Local bandwidth at $\mathbf{x}_{ji}$ is given by $h_{j(i)} = h_{\circ j}\{G/\hat{f}_{jh_{\circ j}}(\mathbf{x}_{ji})\}^{0.5}$, where $G = \{\prod_{i=1}^{n_j} \hat{f}_{jh_{\circ j}}(\mathbf{x}_{ji})\}^{1/n_j}$, and the adaptive kernel density estimate is obtained as $\hat{f}_j^A(\mathbf{x}) = n_j^{-1} \sum_{i=1}^{n_j} h_{j(i)}^{-d} K\left\{h_{j(i)}^{-1}(\mathbf{x} - \mathbf{x}_{ji})\right\}$. Justification for using such locally adaptive bandwidth is given in Abramson (1982) and also in Silverman (1986). To differentiate this classifier from our proposed adaptive classification technique, we will refer to it as the classifier with adaptive kernel density estimates (AKDE).

We start with four well known data sets namely the synthetic data, the vowel data, the image segmentation data and the sonar data. Each of these data sets has specific training and test sets. We

took the synthetic data from CMU data archive (http://www.statlib.cmu.edu), while the other three data sets were taken from UCI machine learning repository (http://www.ics.uci.edu/∼mlearn). As we have mentioned before, in each case, we standardized the data set using usual moment based estimate of the pooled dispersion matrix before using different methods for classification. Recall that our proposed method requires the probability function P involved in ψ_h^+ to be estimated from the training data set to find out the adaptive bandwidth for classification (see Section 3). For this purpose we used both, the estimate based on bootstrap technique and the estimate based on normal approximation as discussed in Section 3, and reported the misclassification rates for both of them (see Table 4.1). For all these data sets, proportions of different classes in the training sample are used as their prior probabilities.

Sonar data : This data set contains 111 patterns obtained by bouncing sonar signals off a metal cylinder and 97 patterns obtained from cylindrical rocks at various angles and under various conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. 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. This data set was split into a training set and a test set of size 104 each using a cluster analysis method to ensure even matching (see Gorman and Sejnowski, 1988). There were 49 patterns from metal cylinder and 55 patterns from rock in the training data set, while the test set consisted of 62 and 42 patterns, respectively, from these two classes.

In this data set, apart from the optimal bandwidth classifier (which had an error rate of 13.42%), the rest of the kernel density estimate based classifiers led to the same test set error rate of 11.9%. All these classifiers misclassified 14 out of 62 test set observations from class-1 (metal cylinder) and 1 out of 42 test set observations from class-2 (cylindrical rock). For this data set, misclassification rates of different classification methods are given in Ripley (1994) and Cooley

and MacEachern (1998). The performance of our proposed method is better than most of these reported results.

Image segmentation data : This data set contains 19 different measurements on each 3×3 pixel image of one of seven different objects : brickface, sky, foliage, cement, window, path and grass. There are 210 observations in the training set and 2100 observations in the test set which are equally distributed among those 7 classes. This data set and descriptions of the measurement variables are available at UCI machine learning repository. The value of the measurement variable 'region pixel count' is '9' for all observations. For the two variables, 'short line density-5' and 'short line density-2', almost 95% of the values are zero. We did not consider these variables in our study. There are some variables in the data set which are linear or nonlinear functions of R ('raw red mean'), B ('raw blue mean') and G ('raw green mean'). We have deleted those variables too and carried out our analysis using the remaining 9 variables.

In this data set, our proposed classifier clearly outperformed its competitors. Misclassification rate of the MISE bandwidth classifier (error rate = 14.71%) and that of the classifier with AKDE (error rate = 14.76%) were more than twice the error rates of our proposed method. The optimal bandwidth classifier performed somewhat better but its error rate (10.38%) was much higher than that of our proposed method (error rate = 6.48% for normal approximation, 6.95% for bootstrap method).

Synthetic data : This data set consists of observations from two different classes, each of which is an equal mixture of two bivariate normal distributions, which differ only in their location parameters. These parameters were chosen to have a Bayes risk of 8.0%. There are 250 observations in the training set and 1000 observations in the test set, which are equally distributed in these two classes. A scatter plot of this data set is given in Figure 4.1, where the dots and the crosses represent the observations from the two classes.

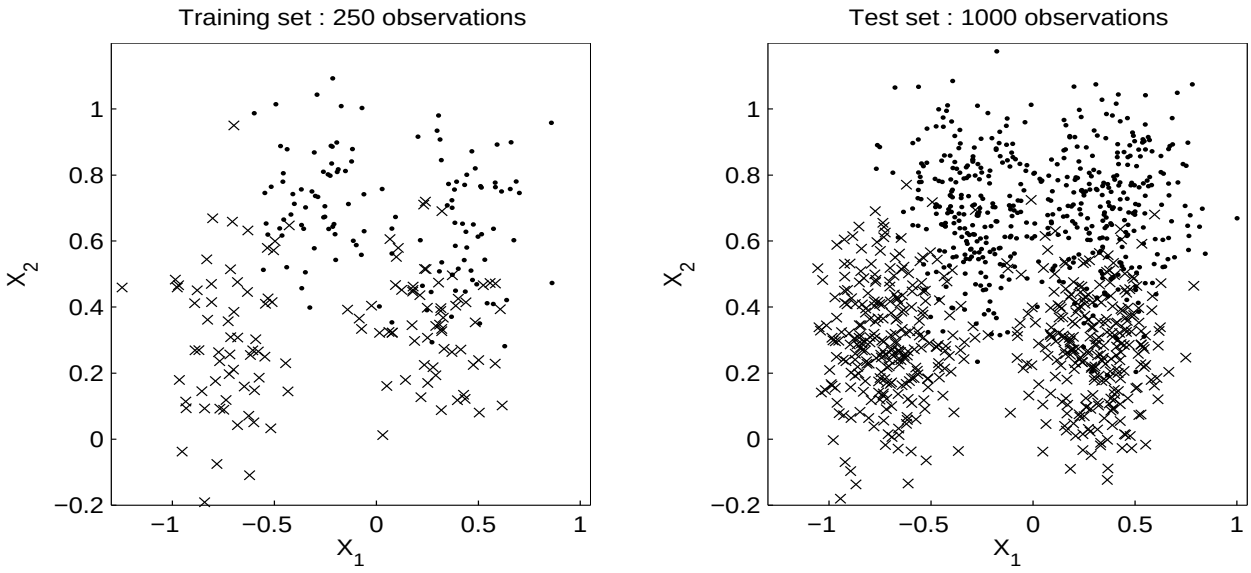

Figure 4.1 : Scatter plot for synthetic data.

In this data set, all kernel density estimate based classifiers that we have considered in this article led to the same misclassification rate of 9.4%. Ripley (1994) used this data set to compare the performance of different classification algorithms. The performance of our proposed method is quite comparable to those reported results.

Vowel data : This data set is related to a vowel recognition problem. Ten different measurements were taken on a speech signal when 90 different speakers spoke 11 different words, each starting with 'h' followed by one of the 11 vowels and then followed by 'd' [see Robinson (1989) for detail description of this data set]. These 990 observations were split into a training set consisting of 528 observations and a test set consisting of 462 observations, which were equally distributed among the 11 classes.

This is a difficult data set and a part of this difficulty arises due to the presence of a fairly large number of overlapping classes. This data set was also used by Hastie, Tibshirani and Buja (1994) and Hastie, Tibshirani and Friedman (2001), where the authors reported the error rates for different classifiers. On this data set, the optimal bandwidth classifier and the proposed method performed better than the MISE bandwidth classifier and the classifier with AKDE. However, the

optimal bandwidth classifier had a slight edge over the proposed method.

Data sets	MISE band. classifier	Classifier with AKDE	Optimal band. classifier	Proposed method	
				Normal app.	Bootstrap
Sonar	11.90	11.90	13.42	11.90	11.90
Image	14.71	14.76	10.38	6.48	6.95
Synthetic	9.40	9.40	9.40	9.40	9.40
Vowel	66.23	66.45	44.59	47.83	47.62

Table 4.1 : Average misclassification rates (in %) of different classifiers.

It is to be noted that in all these data sets, both versions of our proposed method (i.e. the method based on bootstrap estimate and the method based on normal approximation) led to similar misclassification rates. This result is quite encouraging because in the presence of large data sets, due to computational difficulty, it becomes almost impossible to adopt the bootstrap technique. For instance in the case of synthetic data, the method based on bootstrap technique took more than 10 hours on a pentium-4 machine to classify 1000 test sets observations, when 1000 bootstrap samples were used to find the adaptive bandwidths over a grid of 200 points. The method based on normal approximation in this case took only two and half minutes to complete this task when adaptive bandwidths were searched over the same search space of 200 grid points.

Now, we use four other data sets, namely the biomedical data, the chemical and overt diabetes data, the BUPA liver disorder data and the salmon data, for further illustration of our proposed method. Unlike the previous data sets, these new data sets do not have any specific training and test sets. We formed these training and test sets by randomly partitioning the data in such a way that the proportions of different classes in the training and the test sets are as close as possible. This random partitioning was carried out 250 times to generate 250 different training and test sets. Average test set misclassification rates of different methods over these 250 partitions are reported in Table 4.2 along with their corresponding standard errors. Sizes of the training and test

samples in each partition are also reported in the table. We took the BUPA liver disorder data from UCI machine learning repository (http://www.ics.uci.edu/∼mlearn), the biomedical data and the chemical and overt diabetes data from CMU data archive (http://www.statlib.cmu.edu) and the salmon data from the book by Johnson and Wichern (1992). In all these cases, the data sets were standardized using the usual moment based estimate of the pooled dispersion matrix as before, and the sample proportions of different classes are used as their prior probabilities. Because of computational difficulty in reporting the error rates over repeated partitions, we could not use the bootstrap method. Therefore the reported results for our proposed method are based on normal approximation of the distributions of the kernel density estimates as discussed in Section 3.

Biomedical data : This data set was used by Cox, Johnson and Kafadar (1982) in the annual meeting on "Exposition of Statistical Graphics Technology" sponsored by the American Statistical Association Committee on Statistical Graphics. It contains information on four different measurements on each of 209 blood samples taken from normal people as well as from carriers of a genetic disorder. Out of these 209 observations, 15 have some missing values. We removed those 15 observations from the data set and carried out our analysis with remaining 194 observations (127 from normal people and 67 from carriers). On this data set, the MISE bandwidth classifier, the optimal bandwidth classifier and the classifier with AKDE achieved almost similar misclassification rates, but as compared to these methods, our proposed classifier led to much improved performance (see Table 4.2). As compared to the standard errors of average misclassification rates reported inside the braces, this improvement over the competing methods was found to be statistically significant.

Chemical and overt diabetes data : It contains information on five measurement variables (fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight) on blood samples taken from 145 individuals during a three hour glucose tolerance test following an oral administration of glucose. Blood samples were taken from overt diabetic patients, chemical diabetic patients and normal individuals. According to some clinical

classification, there were 33, 36 and 76 observations from these three classes. Reaven and Miller (1979) used different clustering algorithms on this data set to form three clusters and to compare the performance of the clustering methods to the classification results obtained by medical criteria. The latter classification for each patient is given in this data set. Like biomedical data, in this data set our proposed method outperformed its competitors (see Table 4.2). Its performance was significantly better than the other three methods reported in this article. However, there was no significant difference among the error rates of the other three classifiers.

BUPA liver disorder data : This data set was created by BUPA Medical Research Ltd. Here, each observation contains a record of blood test of a single male individual. There are five measurement variables, which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. We did not consider the sixth variable 'number of half-pint equivalents of alcoholic beverages drunk per day' for our analysis. Some sort of selector field was used on this data set to split it into two sets representing the two classes. More description on this data is available at UCI machine learning repository. On this data set the optimum bandwidth classifier and our proposed method performed better than the other two classifiers. The difference between the error rate of the optimum bandwidth classifier and that of the proposed method was found to be statistically insignificant.

Salmon data : This data set consists of 100 bivariate observations on growth ring diameter (freshwater and marine water) of salmon fish coming from Alaskan or Canadian water (50 from each population). A scatter plot of this data set is given in Figure 4.2, where dots and crosses represent the observations coming from Alaskan and Canadian populations, respectively.

It is quite well known that the distribution of the two classes in this data set are nearly elliptic, and linear discriminant analysis performs quite well. We also know that if a large bandwidth is used for kernel discriminant analysis and if the prior probabilities of the competing populations are equal, the resulting classifier behaves like a linear classifier (see e.g., Scott, 1992; Ghosh and

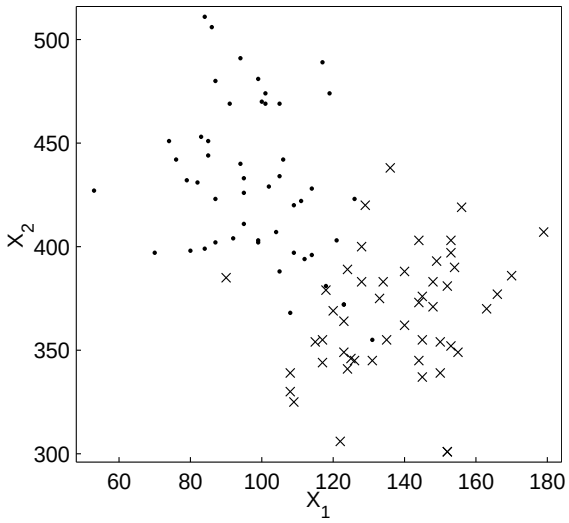

Figure 4.2 : Scatter plot for salmon data.

Chaudhuri, 2004). Therefore, in this data set, it is quite possible to achieve a low misclassification rate using a fixed but large bandwidth for all observations. That is why we observed that the optimum bandwidth classifier performed better than its competitors. However, the proposed method could achieve reasonable misclassification rate.

Data sets	Sample size		MISE band. classifier	Classifier with AKDE	Optimum band. classifier	Proposed method
	Train	Test				
Biomed	100	94	16.76 (0.22)	16.48 (0.20)	16.61 (0.21)	15.41 (0.18)
Diabetes	100	45	12.28 (0.27)	12.23 (0.27)	12.20 (0.28)	11.30 (0.26)
BUPA	200	145	34.06 (0.19)	34.17 (0.19)	33.70 (0.19)	33.67 (0.18)
Salmon	50	50	8.20 (0.19)	8.34 (0.20)	7.38 (0.19)	8.06 (0.18)

Table 4.2 : Average misclassification rates (in %) of different classifiers and their standard errors.

Throughout this article, for a given $\mathbf{x}$, though we have used the same bandwidth for all populations, in practice, one may use J different bandwidths $h_1, h_2, \ldots, h_J$ for J populations. However, it is our empirical experience that use of different bandwidths for different populations generally leads to substantial increase in the computational cost but not necessarily better misclassification

rates. To get an intuitive feeling about this please note that if the ratio h_1/h_2 is very different from 1, even in the degenerate case $f_1 = f_2$, the difference between the expected values of $\hat{f}_{1h_1}$ and $\hat{f}_{2h_2}$ can be significant, and this results in overall classification performance which is less optimal. On the other hand, if we use t grid points for each bandwidth parameter, for a given $\mathbf{x}$, finding the adaptive bandwidth requires a search over t^J points, which grows up exponentially with the number of classes. So, for the computational simplicity, in this article, for a given $\mathbf{x}$, we have used the same bandwidth for all classes after standardization. Nevertheless, from the description of our methodology, it is easy to notice that the proposed adaptive bandwidth selection technique can be adopted for multiple bandwidth problems as well.

5 Concluding remarks

This article presents an adaptive classification technique using kernel density estimates. Instead of fixing the values of the bandwidth parameters, it chooses the bandwidth depending on the observation to be classified and thereby makes the method more flexible. In practice, use of common bandwidth over the whole measurement space may not be a good choice, especially when the true class boundaries are smooth in one part of the measurement space and wiggly in other. Use of spatially adaptive choice of bandwidth is helpful in such situation. Based on the analysis of several benchmark data sets, it may be concluded that the use of adaptive choice of bandwidth generally leads to better estimates of class boundaries and hence improves the performance of the resulting kernel classifier.

Acknowledgement

We are thankful to an associate editor a referee for their careful reading of an earlier version of the paper and for providing us with several helpful comments.

Appendix

Proof of Theorem 2.1 : Since the kernel density estimate $\hat{f}_{jh}(\mathbf{x})$ is an average of iid random variables, its expectation and the variance can be written as

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

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

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

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

$$\begin{aligned}
E_{f_j}\left[K\{(\mathbf{x}-\mathbf{X})/h)\}\right] &= K(\mathbf{0}) + (1/2h^2)E_{f_j}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} + O(h^{-3}) \text{ and} \\
Var_{f_j}\left[K\{(\mathbf{x}-\mathbf{X})/h)\}\right] &= Var_{f_j}\left[(1/2h^2)\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} + (1/6h^3)\sum_{i,j,k} Y_{i,j,k}\right] \\
&= (1/4h^4)Var_{f_j}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} + O(h^{-5}),
\end{aligned}$$

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

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

From the above expression it is quite transparent that for every $j = 1, 2, \ldots, J$, as the sample size n_j tends to infinity, the variance of the kernel density estimate $\hat{f}_{jh}(\mathbf{x})$ tends to zero, and therefore $\hat{f}_{jh}(\mathbf{x})$ converges to its mean $\mu_{jh}(\mathbf{x}) = E\left\{\hat{f}_{jh}(\mathbf{x})\right\}$ in probability. Hence, for all $j \neq i$, the probability function $P\{\pi_j \hat{f}_{jh}(\mathbf{x}) > \pi_i \hat{f}_{ih}(\mathbf{x})\}$ either converges to 1 or to 0 depending on the values of $\pi_j E\{\hat{f}_{jh}(\mathbf{x})\}$ and $\pi_i E\{\hat{f}_{ih}(\mathbf{x})\}$.

Fact 1 : Now, consider the case $\pi_j > \pi_i$. It is quite clear from the expression of the mean and the variance of $\hat{f}_{jh}$ that under this condition, whatever may be the value of $\mathbf{x}$, for large h, $\pi_j \mu_{jh}(\mathbf{x}) = E\{\pi_j \hat{f}_{jh}(\mathbf{x})\}$ will be greater than $\pi_i \mu_{ih}(\mathbf{x}) = E\{\pi_i \hat{f}_{ih}(\mathbf{x})\}$ and hence the probability

function $P\{\pi_j \hat{f}_{jh}(\mathbf{x}) > \pi_i \hat{f}_{ih}(\mathbf{x})\}$ will converge to 1. Similarly, when $\pi_j < \pi_i$, this probability will converge to 0.

Fact 2 : When π_j's are all equal, for large h the ordering of $\pi_j \mu_{jh}(\mathbf{x})$ depends on that of $E_{f_j}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} = E_{f_j}\{(\mathbf{x}-\mathbf{X})' A(\mathbf{x}-\mathbf{X})\}$. Now, notice that

$$\begin{aligned} E_{f_j}\{(\mathbf{x}-\mathbf{X})' A(\mathbf{x}-\mathbf{X})\} &= E_{f_j}[\{(\mathbf{x}-\boldsymbol{\mu}_j)-(\mathbf{X}-\boldsymbol{\mu}_j)\}' A[(\mathbf{x}-\boldsymbol{\mu}_j)-(\mathbf{X}-\boldsymbol{\mu}_j)\}] \\ &= (\mathbf{x}-\boldsymbol{\mu}_j)' A(\mathbf{x}-\boldsymbol{\mu}_j) + E_{f_j}\{(\mathbf{X}-\boldsymbol{\mu}_j)' A(\mathbf{X}-\boldsymbol{\mu}_j)\} \text{ [since } E_{f_j}(\mathbf{X}-\boldsymbol{\mu}_j)=0] \\ &= (\mathbf{x}-\boldsymbol{\mu}_j)' A(\mathbf{x}-\boldsymbol{\mu}_j) + trace[A E_{f_j}\{(\mathbf{X}-\boldsymbol{\mu}_j)(\mathbf{X}-\boldsymbol{\mu}_j)'\}] \\ &= \mathbf{x}' A\mathbf{x} + \boldsymbol{\mu}_j' A \boldsymbol{\mu}_j' - 2\mathbf{x}' A\boldsymbol{\mu}_j + trace(A\boldsymbol{\Sigma}_j) \end{aligned}$$

The first term $\mathbf{x}' A\mathbf{x}$ is independent of the population distribution. So, the ordering of $\pi_j \boldsymbol{\mu}_{jh}(\mathbf{x})$ depends on that of $\boldsymbol{\mu}_j' A \boldsymbol{\mu}_j' - 2\mathbf{x}' A\boldsymbol{\mu}_j + trace(A\boldsymbol{\Sigma}_j)$.

Fact 3 : Consider the following inequality

$$\begin{aligned} \sum_{i\neq j} P\{\pi_j \hat{f}_{jh}(\mathbf{x}) > \pi_i \hat{f}_{ih}(\mathbf{x})\} - (J-2) &\leq P\{\pi_j \hat{f}_{jh}(\mathbf{x}) > \pi_i \hat{f}_{ih}(\mathbf{x}) \text{ for all } i \neq j\} \\ &\leq \min_{i\neq j} P\{\pi_j \hat{f}_{jh}(\mathbf{x}) > \pi_i \hat{f}_{ih}(\mathbf{x})\}. \end{aligned}$$

It is easy to notice that $P\{\pi_j \hat{f}_{jh}(\mathbf{x}) > \pi_i \hat{f}_{ih}(\mathbf{x}) \text{ for all } i \neq j\}$ converges to 1 if and only if for all $i \neq j$, $P\{\pi_j \hat{f}_{jh}(\mathbf{x}) > \pi_i \hat{f}_{ih}(\mathbf{x})\}$ converges to 1. Otherwise, it tends to 0.

Now, as a consequence of the above facts, the rest of the proof follows immediately.

References

[1] Abramson, I. (1982) On bandwidth variation in kernel estimates - a square root law. *Ann. Statist.*, **10**, 1217-1223.

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

[3] Breiman, L., Miesel, W. and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135-144.

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

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

[6] Cooley, C.A. and MacEachern, S. N. (1998). Classification via Kernel Product Estimators. *Biometrika* **85**, 823-833.

[7] Cox, L. H., Johnson, M. M. and Kafadar, K. (1982) Exposition of statistical graphics technology. *ASA Proceedings of the Statistical Computation Section*, pp. 55-56.

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

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

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

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

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

[13] Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2006) Classification using kernel density estimates : multi-scale analysis and visualization. *Technometrics*, **48**, 120-132.

[14] 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.

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

[16] Hall, P. and Wand, M. P. (1988). On nonparametric discrimination using density differences. *Biometrika* **68**, 287-294.

[17] 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.

[18] Hall, P. and Kang, K-H. (2005) Bandwidth choice for nonparametric classification. *Ann. Statist.*, **33**, 284-306.

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

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

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

[22] Johnson, R. A. and Wichern, D. W. (1992) *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey.

[23] 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.

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

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

[26] Muller, H. G. (1984). Smooth optimum kernel estimators of densities, regression crves and modes. *Ann. Statist.* **12**, 76-774.

[27] 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.

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

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

[30] Robinson, A. (1989) Dynamic error propagation networks. *Ph.D. Thesis, Electrical Engineering Department, Cambridge University*.

[31] Sain, S. R. and Scott, D. W. (1996). On locally adaptive density estimation. *J. Amer. Statist. Assoc.*, **91**, 1525-1534.

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

[33] 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.

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

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

[36] Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, **64**, 29-35.

[37] Terrel, G. and Scott, D. W. (1992). Variable kernel density estimation. *Ann. Statist.*, **20**, 1236-1265.

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