SOME NONPARAMETRIC AND SEMIPARAMETRIC METHODS FOR DISCRIMINANT ANALYSIS

ANIL KUMAR GHOSH

Thesis Submitted to the Indian Statistical Institute
in Partial Fulfilment of the Requirements for the
Award of the Degree of **Doctor of Philosophy**

INDIAN STATISTICAL INSTITUTE
203, BARRACKPORE TRUNK ROAD
CALCUTTA-700108, INDIA.

APRIL 2004

Acknowledgement

I take this opportunity to express my sincere gratitude to those persons without whose kind co-operation it would not be possible for me to complete my Ph.D. thesis. The name that comes first in this list is Prof. Probal Chaudhuri, my Ph.D. supervisor. Any expression of gratitude towards him will be an understatement. When I joined this institute as a research fellow, I had some scattered knowledge in different fields of statistics. Constant guidance of Prof. Chaudhuri molded them in a regular shape. His honest criticism in all aspects of my research helped me immensely to improve my research work. In the midst of his endless national and international commitments, he squeezed out sufficient times for me to clarify my innumerable doubts and queries. This thesis is the result of his enormous support and constant inspiration which he provided me throughout my research. The contents of most of the chapters of this thesis appeared as joint work with Prof. Chaudhuri, and he kindly gave me the permission to include them in my thesis.

In the days of my research, I had the opportunity to discuss with Prof. Debasis Sengupta and Prof. C. A. Murthy, whose critical comments and suggestions helped me to improve my research work substantially. The materials of Chapter 3 of this thesis appeared as joint work with Prof. Sengupta who kindly permitted me to include them in my thesis. The contents of a part of Chapter 4 appeared as a joint work with Prof. Murthy and he also gave me the permission to include them.

I am indebted to Dr. Smarajit Bose, who introduced me to the field of nonparametric discriminant analysis. For the first few years of my research, when I had very little idea in this area, he spent his valuable times to make me familiar with this field of research. I am grateful to all teachers of this institute who taught me during my M.Stat and research courses. I am also thankful to my teachers in Presidency College and my tutor Mr. Saroj Mullick, who introduced me to the subject "Statistics". During my M.Stat course, I had the opportunity to share my knowledge with my seniors like Dr. Tirthankar Bhattacharyya and Dr. Sukanta Pati from whom I got the initial inspiration of doing research. I owe them a lot. I am thankful to the members of Research Fellow Advisory Committee and all members of Stat-Math Unit who gave me the much needed moral support. The same is true for my

senior Dr. Saurabh Ghosh and my friends Analabha, Subhadip, Sabyasachi, Sasthi, Partha, Tapas, Rajib, Sujit and others. I am also grateful to Prof. Alok Goswami, Dr. Debashis Goswami and Mr. Gautam Das for useful academic discussions. Nonacademic staff of the Stat-Math Unit, the Reprography Unit, the Dean's office and the library helped me in various phases of my tenure of as a research fellow. I am in their debt. I am also thankful to Dr. Sumitra Purakayastha and Mr. Swarup De for providing me better computing facilities, and to Mr. Dibyendu Bose for helping me in preparing the LaTex print out of this thesis. I should also acknowledge my juniors in this department and outside who provided me entertaining relief during my period of monotonicity.

Finally, I express my gratitude to my parents, sister and other members of my family, specially my uncle Mr. Arun Ghosh and cousin Mr. Gour Ghosh, who encouraged me through out my research work. I am also thankful to my wife Rachana, who was the constant source of my inspiration right from the beginning of my research. But this list will remain incomplete if I forget to mention my little nieces Anindita, Arpita and Oindrilla who stretched out their little hands as much as possible to help me in preparing this thesis.

April, 2004
Calcutta.

Anil Kumar Ghosh

Contents

List of Tables

List of Figures

Figures **Page**

Notations and Abbreviations

$\mathbf{x}, \mathbf{X}$	: measurement vectors
J	: number of populations/classes
d	: dimension of the measurement vector
N	: size of the training smaple
Δ	: misclassification rate
$\mathbf{d}(\mathbf{x})$	: decision rule
$\mathbf{d}_B(\mathbf{x})$	: Bayes classification rule
$\mathbf{d}_K(\mathbf{x})$	: kernel classifier
$\mathbf{d}_l(\mathbf{x})$, $\mathbf{d}_L(\mathbf{x})$	: linear classifiers
$\mathbf{d}_Q(\mathbf{x})$	: quadratic classifier
$\mathbf{d}_D(\mathbf{x})$	: maximum depth classifier
$\mathbf{d}_{D^*}(\mathbf{x})$	: maximum depth classifier modified for unequal prior cases
μ	: location parameter
Σ	: scatter matrix
$\mathbf{I}_d$	: identity matrix of the order $d \times d$
$\mathbf{x}_k$	: k^{th} observation of the training sample
c_k	: class label of $\mathbf{x}_k$
$f(\mathbf{x})$	: density function of $\mathbf{x}$
$\hat{f}(\mathbf{x})$	: density estimate of $f(\mathbf{x})$
$K(\mathbf{t})$	: kernel density function at $\mathbf{t}$
π_j	: prior probability of the j^{th} population
μ_j	: location parameter of the j^{th} population
Σ_j	: scatter matrix of the j^{th} population
$\mathbf{x}_{jk}$	: k^{th} observation of the j^{th} population
$\Phi(x)$	: c.d.f. of standard normal distribution at x
$\phi(x)$	: p.d.f. of standard normal distribution at x
$\phi(x, \mu, s)$	: p.d.f. of a normal distribution (with mean μ and variance s^2) at x
$\phi_d(\mathbf{x}, \mu, \Sigma)$	: p.d.f. of a d-dimensional multivariate normal distribution (with mean μ and dispersion matrix Σ) at $\mathbf{x}$

n_j : number of training sample observations in the j^{th} class

$\mathbf{n}$: $(n_1, n_2, \ldots, n_J)$, vector of training sample sizes

$p(j \mid \mathbf{x})$: posterior probability of the j^{th} population given the observation $\mathbf{x}$

h : bandwidth parameter

$f_j(\mathbf{x})$: density function of $\mathbf{x}$ in the j^{th} population

$\hat{f}_{jh}(\mathbf{x})$: kernel density estimate for $f_j(\mathbf{x})$ when h is used as the bandwidth parameter

$\Delta(h)$: misclassification rate of the kernel classifier where h is used as the common bandwidth for all competing populations

h_o : optimal bandwidth for density estimation (bandwidth that minimizes mean integrated square error of the kernel density estimate)

h_* : optimal bandwidth for classification

h^+ : largest bandwidth leading to the lowest leave-one-out cross-validation error

$h^+_{(10)}$: largest bandwidth leading to the lowest 10-fold cross-validation error

h_j : bandwidth parameter for the j^{th} population

$\mu_{jh_j}(\mathbf{x})$: expectation of $\hat{f}_{jh_j}(\mathbf{x})$

$s^2_{jhj}(\mathbf{x})$: variance of $\hat{f}_{jh_j}(\mathbf{x})$

$\mu^*_{jh_j}(\mathbf{x})$: data based estimate for $\mu_{jh_j}(\mathbf{x})$

$s^{*2}_{jh_j}(\mathbf{x})$: data based estimate for $s^2_{jhj}(\mathbf{x})$

$\Delta(h_1, h_2, \ldots, h_J)$: misclassification rate of the kernel classifier when $h_1, h_2, \ldots, h_J$ are used as the bandwidth parameters for different classes

$\psi(h_1, h_2, \ldots, h_J)$: approximate version of $\Delta(h_1, h_2, \ldots, h_J)$ under normality assumption

$\psi_N(h_1, h_2, \ldots, h_J)$: an empirical analogue of $\psi(h_1, h_2, \ldots, h_J)$

$\psi^*_N(h_1, h_2, \ldots, h_J)$: data based estimate of $\psi_N(h_1, h_2, \ldots, h_J)$

$\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$: posterior probability for population-1 (when it is compared with population-2) given the bandwidth pairs h_1, h_2 and the observation $\mathbf{x}$

$P_{h_1,h_2}(\mathbf{x})$: p-value in favor of population-1 (when it is compared with population-2) given the bandwidth pairs h_1, h_2 and the observation $\mathbf{x}$

R_{h_1,h_2} : collection of all d-dimensional points $\mathbf{x}$ where $\pi_1 \hat{f}_{1h_1}(\mathbf{x}) > \pi_2 \hat{f}_{2h_2}(\mathbf{x})$

$\hat{s}^2_{jhj}(\mathbf{x})$: sample analogue of $s^2_{jhj}(\mathbf{x})$

$\hat{\Delta}$: data based estimate for Δ

Δ_0 : estimated misclassification rate of the best kernel classifier

$w(h_1, h_2)$: weight function for the bandwidth pair (h_1, h_2)

$W_{\mathbf{x}}(h_1, h_2)$: weight function adjusted using the p-value at $\mathbf{x}$

$W^*_{\mathbf{x}}(h_1, h_2)$: sign adjusted version of $W_{\mathbf{x}}(h_1, h_2)$

C_k : k-nearest neighbor classifier

$\mathbf{x}^{(k,N)}$: k-th nearest neighbor of $\mathbf{x}$ in a sample of size N.

$\rho(\mathbf{x}, \mathbf{y})$	: Euclidean distance between two measurement vectors $\mathbf{x}$ and $\mathbf{y}$
r_k	: Euclidean distance between $\mathbf{x}$ and its k^{th} nearest neighbor
$\hat{f}_{j,k_j}(\mathbf{x})$	: nearest neighbor density estimate of $f_j(\mathbf{x})$, when k_j is used as the neighborhood parameter
$V_{j,k_j}(\mathbf{x})$	: volume of the hyper-sphere (with center at $\mathbf{x}$) extending up to k_j^{th} nearest neighbor of $\mathbf{x}$ in the j^{th} population
p_j	: abbreviated form of $p(j \mid \mathbf{x})$ for some given $\mathbf{x}$
$\mathbf{p}$	: $(p_1, p_2, \ldots, p_J)$, vector of p_j's
$\pi(\mathbf{p})$	: prior distribution of $p_1, p_2, \ldots, p_J$
$\hat{p}_j$	: nearest neighbor estimate for p_j
t_{jk}	: number of neighbors (out of k nearest neighbors) from the j^{th} population
$\mathbf{t}_k$	: $(t_{1k}, t_{2k}, \ldots, t_{Jk})$, vector of t_{jk}'s for fixed k
$\varphi(\mathbf{t} \mid \mathbf{p}, k)$	: conditional distribution of $\mathbf{t}$ for given $\mathbf{p}$ and k
$f(\mathbf{p} \mid k, \mathbf{t})$	: conditional distribution of $\mathbf{p}$ for given k and $\mathbf{t}$
$S(j \mid k)$	: Bayesian measure of strength for j^{th} population given the value of k
Δ_o	: estimated error rate of the best nearest neighbor classifier
$\omega(k)$	: weight function for the nearest neighbor classifier C_k
τ	: maximum allowable deviation (in a standard scale) form Δ_0 (or Δ_o) to have positive weight in the weighting scheme $w(h_1, h_2)$ (or $\omega(k)$)
F, G	: distribution functions
Λ	: symmetric matrix
$\sigma(x)$	: sigmoid transformation of x
$\boldsymbol{\alpha}$	: direction vector
$U_{\mathbf{n}}(\boldsymbol{\alpha})$	: linear separation between training samples of two populations along the direction $\boldsymbol{\alpha}$
$U(\boldsymbol{\alpha})$	: population analogue of $U_{\mathbf{n}}(\boldsymbol{\alpha})$
β	: intercept term
$\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$	: training set misclassification error for the classifier $\boldsymbol{\alpha}'\mathbf{x} + \beta = 0$
$\Delta(\boldsymbol{\alpha}, \beta)$	: population analogue of $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$
$\boldsymbol{\alpha}_+$	: augmented vector $(\boldsymbol{\alpha}, \beta)$, a fit
$\hat{\boldsymbol{\alpha}}_H, \hat{\beta}_H$	: estimate of $\boldsymbol{\alpha}$ and β obtained by semiparametric classifier based on half space depth
$\hat{\boldsymbol{\alpha}}_R, \hat{\beta}_R$	: estimate of $\boldsymbol{\alpha}$ and β obtained by semiparametric classifier based on regression depth
$\mathbf{z}$	: transformed vector (projection of $\mathbf{x}$) in a higher dimensional space
r	: dimensionality of the projected space
Ω_n	: data cloud consisting of n observations
$\mathcal{S}$	: discriminating/separating surface
H	: hyperplane

$D(F, \mathbf{x})$	: depth of $\mathbf{x}$ with respect to the distribution F
$D(F_n, \mathbf{x})$	: depth of $\mathbf{x}$ with respect to a data cloud of n observations from F
$D(j, \mathbf{x})$	: population depth of $\mathbf{x}$ in the j^{th} population
$D_{n_j}(j, \mathbf{x})$	: empirical depth of $\mathbf{x}$ in a data cloud of n_j observations from the j^{th} population
$\mathbf{x_0}$	: a future observation
$v_0^{(i)}$	: re-scaled Mahalanobis distance between $\mathbf{x_0}$ and the center of the i^{th} population
$v_{jk}^{(i)}$	: re-scaled Mahalanobis distance between $\mathbf{x}_{jk}$ and the center of the i^{th} population
$\xi_{p,\alpha}$	: p-th population quantile of $\alpha^{'}\mathbf{X}$
$\xi_{p,\alpha,n}$	: empirical version of $\xi_{p,\alpha}$ based on a sample of size n
$\gamma_j\{(D(j, \mathbf{x})\}$	: transformation of $D(j, \mathbf{x})$ that gives the Mahalanobis distance between $\mathbf{x}$ and the center of the j^{th} population
ϱ_j	: density function of the Mahalanobis distance in the j^{th} population.
$\theta_j\{(D(j, \mathbf{x})\}$	: transformation of $D(j, \mathbf{x})$ adopted for maximum depth classification in unequal prior cases
$CART$	: classification and regression trees
CNN	: condensed nearest neighbor
FDA	: flexible discriminant analysis
HD	: half-space depth
LDA	: linear discriminant analysis
$MARS$	: multivariate adaptive regression splines
$MCMC$	: Markov chain Monte carlo technique
$MISE$	: mean integrated square error
MD	: Mahalanobis depth
MJD	: majority depth
PD	: projection depth
QDA	: quadratic discriminant analysis
SD	: simplicial depth
SPD	: spatial depth
SVD	: simplicial volume depth
SVM	: support vector machines

Chapter 1

Introduction

Discriminant analysis (see e.g., Devijver and Kittler, 1982; Duda, Hart and Stork, 2000; Hastie, Tibshirani and Friedman, 2001) deals with the separation of different groups of observations and allocation of a new observation to one of the previously defined groups. In a J-class discriminant analysis problem, we usually have a training sample of the form $\{(\mathbf{x}_k, c_k) : k = 1, 2, \ldots, N\}$, where $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kd})$ is a d-dimensional measurement vector, and $c_k \in \{1, 2, \ldots, J\}$ is its class label. On the basis of this training sample, one aims to form a decision rule $\mathbf{d}(\mathbf{x}) : R^d \to \{1, 2, \ldots, J\}$ for classifying the future observations into one of the J classes with the maximum possible accuracy. The optimal Bayes rule (see e.g., Rao, 1973; Anderson, 1984) assigns an observation to the class which has the largest posterior probability. It can be described as

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

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 ($j = 1, 2, \ldots, J$).

These density functions $f_j(\mathbf{x})$'s are usually unknown in practice, and can be estimated from the training sample either parametrically or nonparametrically. Parametric approaches (see e.g., Rao, 1973; Mardia, Kent and Bibby, 1979; Anderson, 1984; James, 1985; Fukunaga, 1990; McLachlan, 1992) are motivated by some specific distributional assumptions about the underlying populations, where forms of the density functions are assumed to be known except for some unknown real parameters (e.g., means, variances, correlations). For instance, Fisher's linear and quadratic discriminant analysis (Fisher, 1936) are mainly motivated by the normality of the population distributions. Consequently, the performance of these parametric discrimination rules largely depends on the validity of those parametric models. Such model assumptions are usually difficult to verify in practice, and inappropriate models may lead to a rather poor classification. This is why there is a need to develop nonparametric and distribution free methods for discriminant analysis. These nonparamet-

ric classification techniques are more flexible in nature and free from all such parametric model assumptions. Notably, methods like classification trees (see e.g., Breiman *et. al.*, 1984; Loh and Vanichsetakul, 1988; Loh and Shih, 1997; Kim and Loh, 2001, 2003), nearest neighbors (see e.g., Fix and Hodges, 1951; Cover and Hart, 1968; Dasarathy, 1991), flexible discriminant analysis (see e.g., Hastie, Tibshirani and Buja, 1994), splines (see e.g., Bose, 1996; Kooperberg, Bose and Stone, 1997), neural nets (see e.g., Lippman, 1987; Cheng and Titterington, 1994; Ripley, 1994, 1996) and support vector machines (see e.g., Vapnik, 1995, 1998; Burges, 1998) are known to outperform the parametric approaches in a wide variety of problems. A comparative study of the performance of several parametric and nonparametric classification algorithms can be found in the recent paper by Lim, Loh and Shih (2000).

Kernel discriminant analysis (see e.g., Hand, 1982; Coomans and Broeckaert, 1986) is one well-known nonparametric method, which uses the kernel estimates of the population densities (see e.g., Silverman, 1986; Scott, 1992; Wand and Jones, 1995) to construct the classification rule. This application of kernel density estimates in nonparametric discriminant analysis is quite popular in the existing literature, and this popularity is evident in its implementation in some commonly used statistical softwares (e.g., SAS). Like any other nonparametric method, a kernel density estimate involves a smoothing parameter, which is commonly known as the bandwidth. The value of the bandwidth parameter plays an important role both in density estimation and discriminant analysis. Behavior of the kernel density estimate for different choices of the bandwidth parameter is well studied in the literature (see e.g., Silverman, 1986; Scott, 1992; Wand and Jones, 1995), and one usually minimizes the mean integrated square error of the density estimate to find out the optimal bandwidth for density estimation. However, in the existing literature, it has not been properly investigated how the performance of the classification rule is influenced by the values of the bandwidth. In Chapter 2 of this thesis, we make a critical investigation into this problem. Throughout this thesis, the performance of a classification rule is evaluated by its average misclassification probability given by

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

In the course of this above investigation, we observe 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 appropriate regularity conditions. The contents of this chapter are based on Ghosh and Chaudhuri (2004).

The use of cross-validation based methods for bandwidth selection, which try to minimize the estimated average misclassification rate, may require huge computations when there are several competing classes. Besides, such methods usually allow only one bandwidth for each population density estimate, while in a classification problem, the optimum bandwidth for a class density estimate may vary significantly depending on its competing class densities and their prior probabilities. Therefore, in a J-class ($J > 2$) problem, it would be more useful to have different bandwidths for a class density when it is compared with different competing class densities. Moreover, good choice of bandwidths should also depend on the specific observation to be classified. Consequently, instead of concentrating on a single optimum bandwidth for each population density estimate, it would be more useful in practice to look at the results for different scales of smoothing for the kernel density estimates. Chapter 3 of this thesis presents such a multi-scale approach along with a graphical device leading to a more informative discriminant analysis than the usual approach based on a single optimum scale of smoothing for each class density estimate. When there are more than two competing classes, this method splits the problem into a number of two-class problems. This allows the flexibility of using different bandwidths for different pairs of competing classes, and at the same time it reduces the computational burden that one faces for usual cross-validation based bandwidth selection when there are many competing classes. Along with the posterior probability, we use a p-value type measure for discrimination which gives a better graphical representation of the strength of evidence in favor of different competing populations. Uncertainties at various locations of the plot are judged by appropriately estimated misclassification probabilities, and the aggregation of all these informations leads to the final classification. Several benchmark examples are used to illustrate the usefulness of the proposed methodology. The contents of this chapter are based on Ghosh, Chaudhuri and Sengupta (2003a, 2003b).

Like kernel discriminant analysis, nearest neighbor classification is another very popular method of discriminant analysis. Its simplicity has made it a widely used technique in machine learning for pattern recognition problems. One major issue in k-nearest neighbor

classification is to find out an ideal value of the neighborhood parameter k, which is again generally estimated by the method of cross-validation. However, a good choice of k in a classification problem not only depends on the entire training data set but also on the specific observation to be classified. Therefore, like multi-scale kernel discriminant analysis, instead of using a single value of k, it would be more meaningful to use a range of values for k to arrive at the final decision. Simultaneous consideration of different values of k is expected to provide more useful information for classification and its strength than that obtained in a method based on a single value of k. In Chapter 4 of this thesis, we present one such approach, where the results for a finite sequence of classifiers indexed by k are presented in two-dimensional plots along with their corresponding estimated misclassification probabilities. Final decision is obtained by a judicious aggregation of these informations. Along with the usual posterior probability estimates, we use another Bayesian measure for the strength of evidence in favor of different populations. In a two-class problem, this measure is somewhat related to the p-value type measure associated with kernel discriminant analysis described in Chapter 3 of this thesis. When compared with the posterior probability plot, the plot of this Bayesian measure gives an enhanced representation of the difference between the strengths of two competing populations. A similar approach is also adopted for classification using nearest neighbor density estimates. This method allows the flexibility of using different types of neighborhood regions and different values of k for different competing class densities. Like the kernel method, here also it is computationally difficult to minimize the misclassification rate when there are more than two competing populations, and we adopt the pairwise approach to avoid this computational burden. Usefulness of the proposed methodology has been illustrated using some benchmark data sets. The contents of this chapter are mainly drawn from Ghosh, Chaudhuri and Murthy (2003).

In Chapter 5 of this thesis, we focus our attention to some classification procedures that are semiparametric in nature. Though nonparametric classification algorithms are quite flexible in nature and free from all parametric model assumptions, they sometimes have a tendency to overfit the training data. Besides, such methods may require very large number of training set observations to produce good results on test cases. In this chapter, we assume a finite dimensional (linear or nonlinear) parametric form for the discriminating surface. However, instead of estimating the coefficients in the equation of the discriminating surface using specific distributional assumptions (e.g., moment based estimation under the assumption of normal distribution), we use here the idea of data depth to estimate those coefficients so that the resulting classifier has its misclassification rates minimized in some appropriate sense. Over the last couple of decades, data depth has emerged as a powerful exploratory and inferential tool for multivariate data analysis, and it has wide applications in different fields of statistics. Very simple examples of parametric separating surfaces such as a linear or a quadratic classifier tries to find out the linear projection or the quadratic function

of the measurement variables that maximizes the separation between the classes. These techniques are very useful in obtaining good lower dimensional view of class separability. Fisher's (Fisher, 1936) discriminant analysis, which is primarily motivated by the normality of the data distribution, uses the first and the second order moments of the training sample observations to build such classifiers. Consequently, these classifiers are highly sensitive to outliers, and they are not reliable for heavy tailed distributions. In this chapter, we investigate two distribution free methods for linear as well as nonlinear classification, which are based on the notions of statistical depth functions. One of these classifiers uses the idea of Tukey's half-space depth (Tukey, 1975) while the other one is based on the concept of regression depth (Rousseeuw and Hubert, 1999). These depth based procedures use only the distributional geometry of the data cloud for building the classifier. We use a few simulated and real data sets to examine the performance of these discriminant analysis tools and study their asymptotic properties under appropriate regularity conditions. Over a large variety of data sets, these depth based linear and quadratic classifiers have outperformed the traditional linear and quadratic discriminant analysis. The contents of this chapter are taken from Ghosh and Chaudhuri (2003a).

In Chapter 6 of this thesis, we investigate some other types of depth based fully nonparametric classifiers for discriminating between several competing populations. Unlike the classifiers defined in Chapter 5 of this thesis, these methods do not assume any specific parametric form of the separating surfaces and use different notions of data depth to estimate the class boundaries. Data depth provides a center outward ordering of multivariate observations with respect to a multivariate distribution or a given data cloud. Depth of a point is expected to be higher if it lies near the center of the distribution, while it should gradually decrease as the point moves away from the center. The classifiers investigated in this chapter use this idea, and they assign an observation to the population with respect to which it has the maximum location depth. Here, we study different classifiers derived from various depth functions like the half-space depth (Tukey, 1975), simplicial depth (Liu, 1990), majority depth (Singh, 1991), projection depth (Stahel, 1981; Donoho, 1982; Zuo and Serfling, 2000a, 2000b), simplicial volume depth (Oja, 1983; Zuo and Serfling, 2000a, 2000b) and spatial depth (Serfling, 2002). Spatial depth is computationally less expensive than other existing depth functions and gets less affected by the data sparsity in higher dimensions. When prior probabilities of the competing populations are all equal, and the populations are elliptic differing only in their location parameters, we show that these maximum depth classifiers possess certain optimality properties. Possible modifications of these classification rules are also suggested for unequal prior cases. In this chapter, we use some simulated examples and some benchmark data sets to study the performance of these depth based classifiers, and the large sample properties of their error rates have been investigated under appropriate regularity conditions. The contents of this chapter are based on Ghosh

and Chaudhuri (2003b).

In Chapter 7, we indicate some possible directions for further research and some open problems. A complete list of references is given at the end of this thesis.

Chapter 2

Optimal smoothing in kernel discriminant analysis

2.1 Main problem and motivation

Kernel density estimation (see e.g., Muller, 1984; Silverman, 1986; Scott, 1992) is one of the most well known methods 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; Silverman, 1986; Hall and Wand, 1988; Scott, 1992; Bensmail and Bozdogan, 2002) and in many standard softwares (e.g., SAS). If $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are d-dimensional observations in the training sample from the j^{th} population, 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 $K(\cdot)$ is a density function on the d-dimensional space, and $h > 0$ is an associated smoothing parameter popularly known as the bandwidth. A classification rule based on these kernel density estimates can be described as

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

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; Scott, 1992) of the kernel density estimate. 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 $\Sigma = \mathrm{I}$ but different mean vectors μ_1 and μ_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 ($n_1 = n_2 = n = 50$) was good enough for a very high degree of accuracy in the normal approximation for the distribution of kernel density estimates.

In Figure 2.1, values of $\Delta(h)$ has been plotted for varying choices of h and for different dimensions ($d = 1, 2, 4, 6$), where we have chosen $\mu_1 = (0, 0, \cdots, 0)$ and $\mu_2 = (2, 0, \cdots, 0)$, and the sample sizes are taken as 50 for both the 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 'o'. 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 that 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 2.1 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; Wand and Jones, 1995 for detailed discussion). However, in all the cases in Figure 2.1, $\Delta(h)$ becomes almost flat after reaching its minimum value. Unlike what

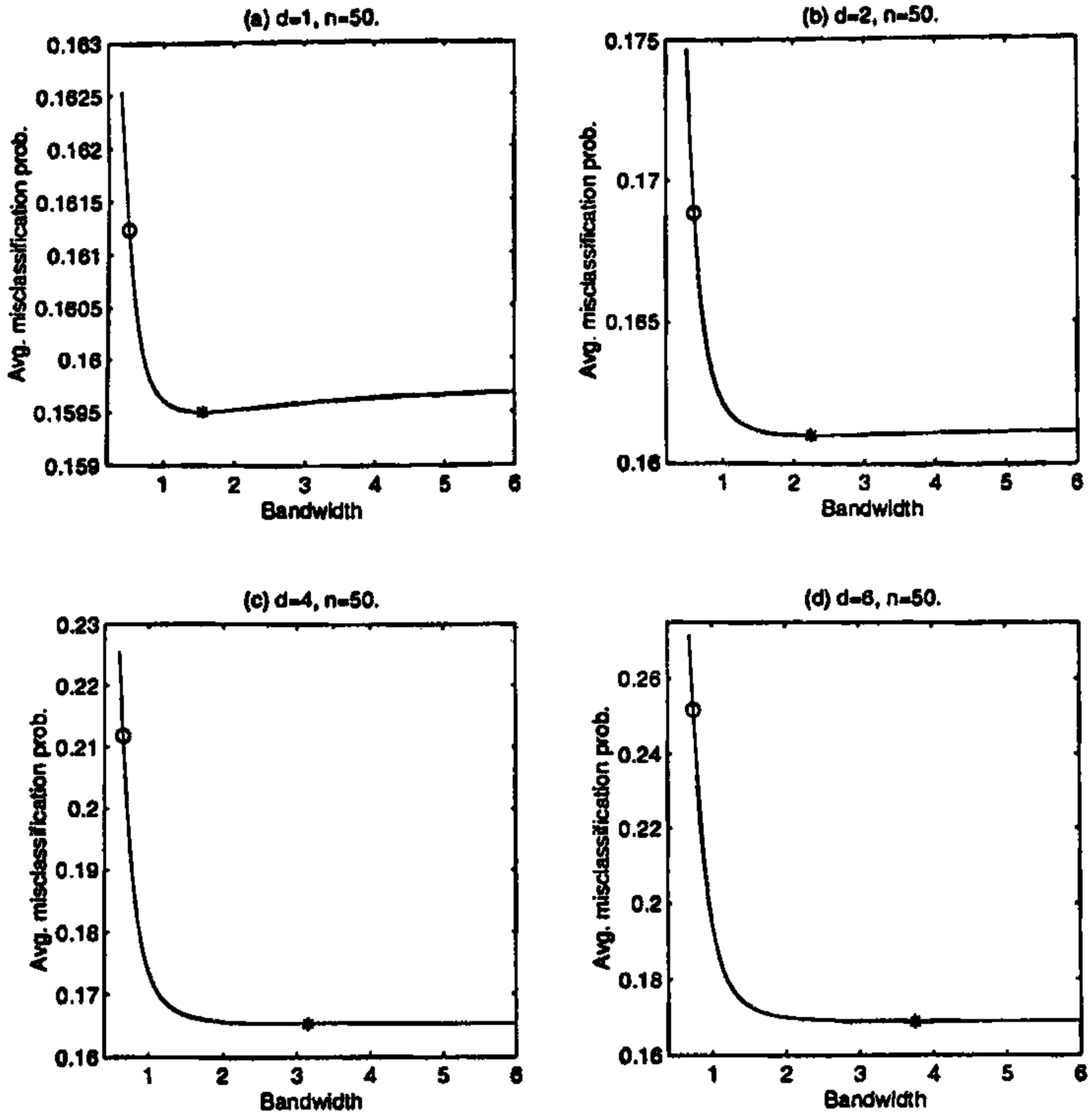


Figure 2.1 : True Δ-functions and optimal bandwidths (equal prior cases)

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 μ_1, μ_2 and Σ, we get different figures for $\Delta(h)$ but the basic pattern remains the same.

There are other popular methods for choosing the bandwidth in a classification problem based on cross-validation techniques (see e.g., Stone [M. Stone], 1977; Ripley, 1996; Duda *et. al.*, 2000). For instance, V-fold cross-validation method divides the whole training sample into V parts of sizes as nearly equal as possible. Usually stratified random sampling is used to form these 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; 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.

Figure 2.2, nicely demonstrates the limitations of such cross-validation based techniques. Here, we have considered the same problem as in Figure 2.1 and generated 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 2.2. The estimated curves not only behave like step functions but also miss the proper locations of optimum bandwidths by wide margins in some cases.

2.2 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), $E\{\hat{f}_{jh}(\mathbf{x})\}$ and $\hat{f}_{jh}(\mathbf{x})$ can be viewed as *the theoretical* and *the empirical scale space functions* of the j^{th} population, 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 $\hat{f}_{jh}(\mathbf{x})$ tends to be almost degenerate at $E\{\hat{f}_{jh}(\mathbf{x})\}$ when the sample size is large.

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 location shift model.

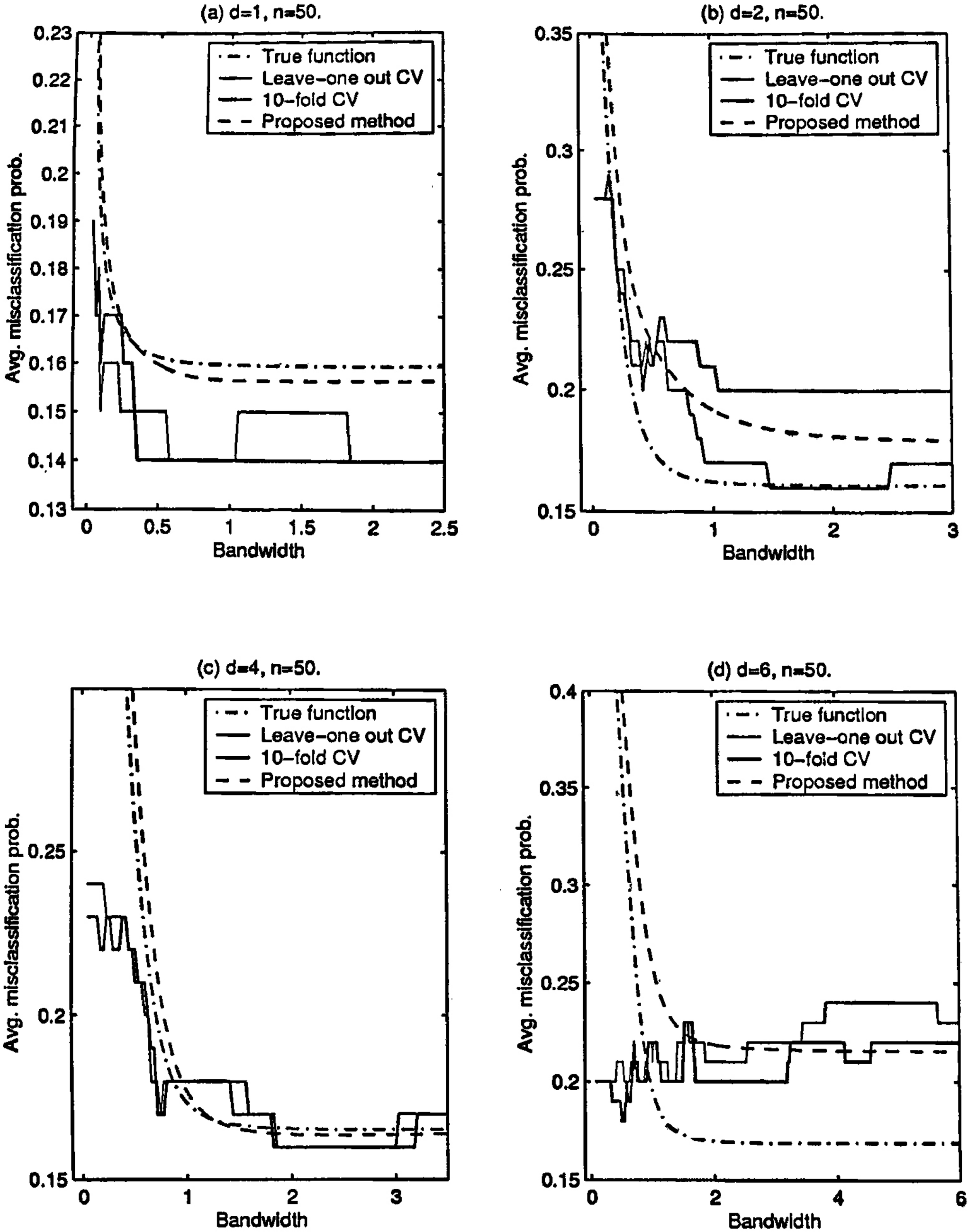


Figure 2.2 : Average misclassification probabilities (equal prior cases)

Theorem 2.1 : *Suppose that $f_1, f_2, \ldots, f_J$ and K are all spherically symmetric densities, and the f_j's satisfy the location shift model i.e., $f_j(\mathbf{x}) = g(\mathbf{x} - \mu_j)$ for some common density g with zero mean and location parameter μ_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 2.1. The next theorem throws some light into the behavior of kernel density estimate based classifiers for large sample sizes and large bandwidths in general when the population densities do not necessarily satisfy any symmetry condition.

Theorem 2.2 : *Suppose that f_j's are density functions satisfying $\int \|\mathbf{x}\|^6 f_j(\mathbf{x})d\mathbf{x} < \infty$ for all $j = 1, 2, \ldots, J$, and the kernel K is a density with a mode at 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*

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

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 $\mathbf{x}' E_{f_j}(\mathbf{X}) - (1/2) E_{f_j}(\mathbf{X}\mathbf{X}')$ (since $\nabla^2 K(\mathbf{0})$ is negative definite) or minimizes $E_{f_j}(\|\mathbf{x} - \mathbf{X}\|^2)$ for $1 \leq j \leq J$.

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 2.1 and 2.2 but now we set the priors at 0.6 and 0.4, respectively, for the two populations. The results obtained are summarized in Figure 2.3. 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 'o') 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.

Theorem 2.3 : *Suppose that the density functions $f_1, f_2, \ldots, f_J$ and the kernel K satisfy the conditions of Theorem 2.2. Assume further that the densities f_j's satisfy*

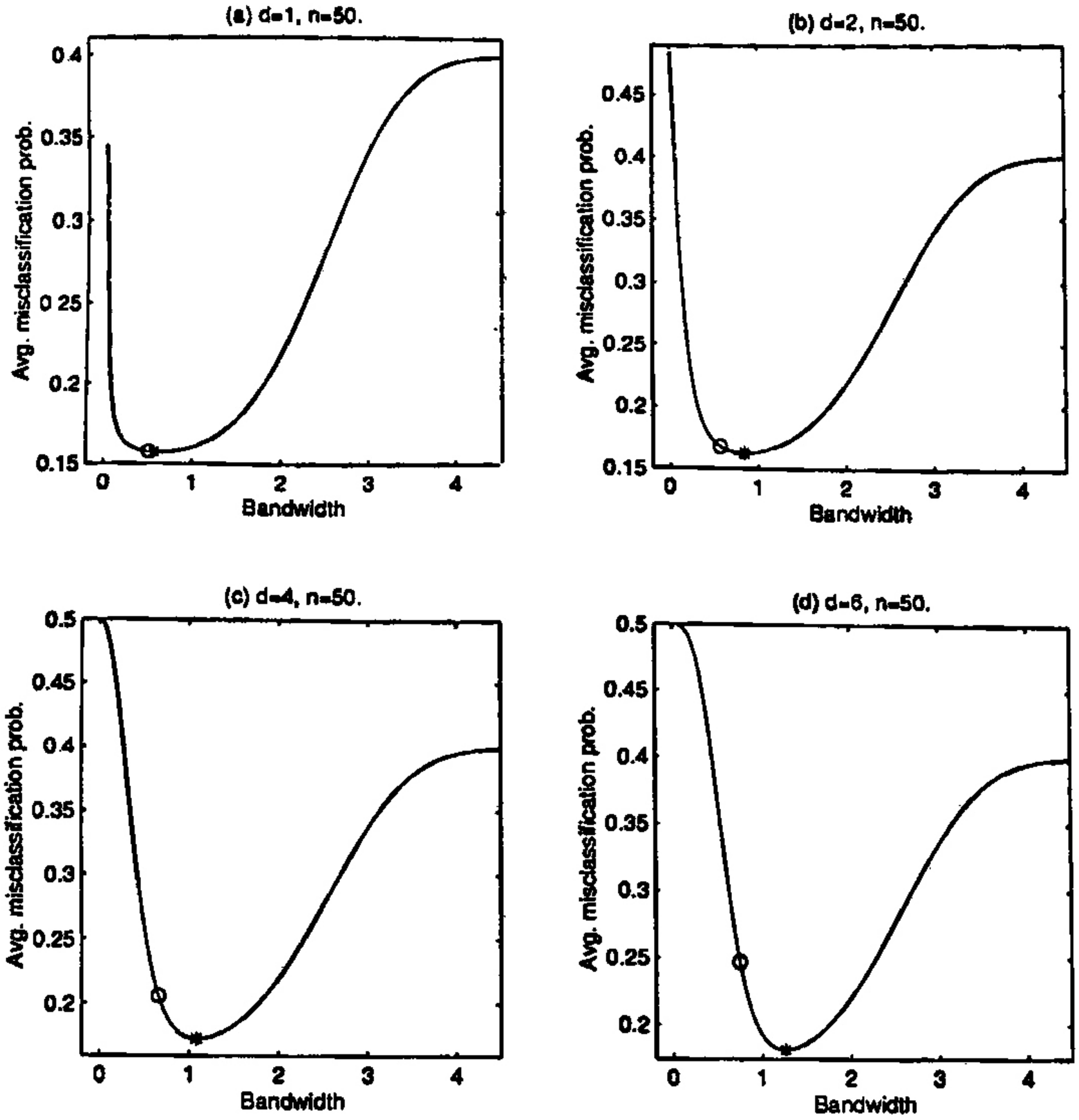


Figure 2.3 : True Δ-functions and optimal bandwidths (unequal prior cases)

the location-shift model in the sense that for all $j = 1, 2, \ldots, J$, $f_j(\mathbf{x}) = g(\mathbf{x} - \mu_j)$ *for a common density* g *with zero mean and location parameter* μ_j. *Then, as* $n, h \to \infty$, *the average misclassification probability of the kernel density estimate based classifier behaves in the following way.*

(*a*) *If* $\pi_1 = \pi_2 = \ldots = \pi_J$, *the average misclassification rate of the classifier tends to that of a linear classifier given by*

$$\mathbf{d}_l(\mathbf{x}) = arg\min_j \left[\mathbf{x}'\nabla^2 K(\mathbf{0})\mu_j - \{\mu_j{}'\nabla^2 K(\mathbf{0})\mu_j\}/2\right].$$

(*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,* $\pi_{j_1} = \pi_{j_2} = \ldots = \pi_{j_m} > \pi_j$ *for all* $j \notin \{j_1, j_2, \ldots, j_m\}$, *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 2.4 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 methods also has 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 usual linear classifier in this case) when the prior probabilities are equal. In Figure 2.5, we have plotted the class 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 (see e.g., Hall, 1983; Stone, 1984; Silverman, 1986; Hall *et. al.*, 1991; Sheather and Jones, 1991; Wand and Jones, 1995) for choosing the bandwidth from the data. 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 2.6, we show the performance of such cross-validatory techniques for the normal population problems with unequal priors (as in Figure 2.3). 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.

2.3 Data based choice for bandwidths

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

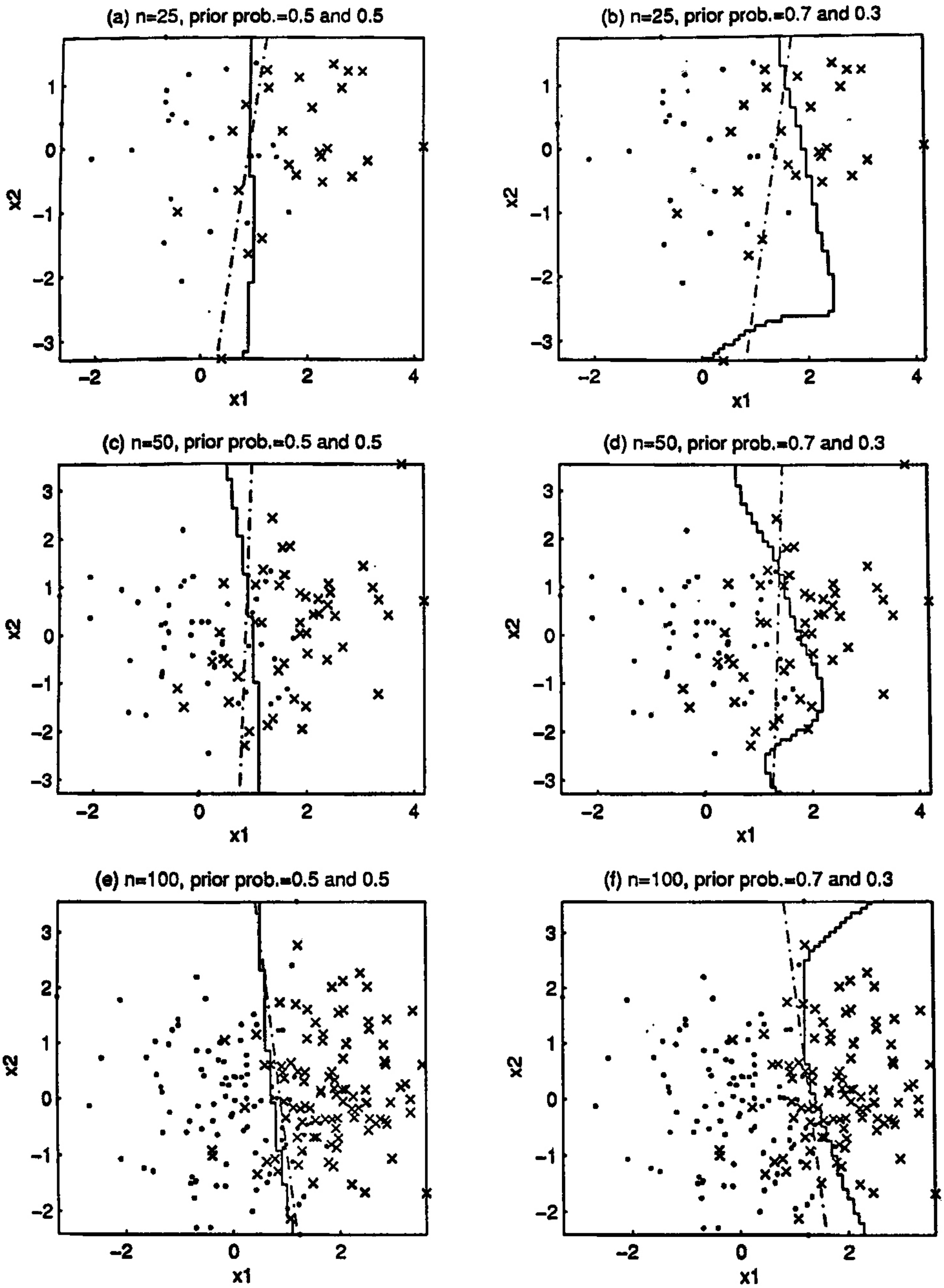


Figure 2.4 : Classification boundaries for kernel based and linear classifiers

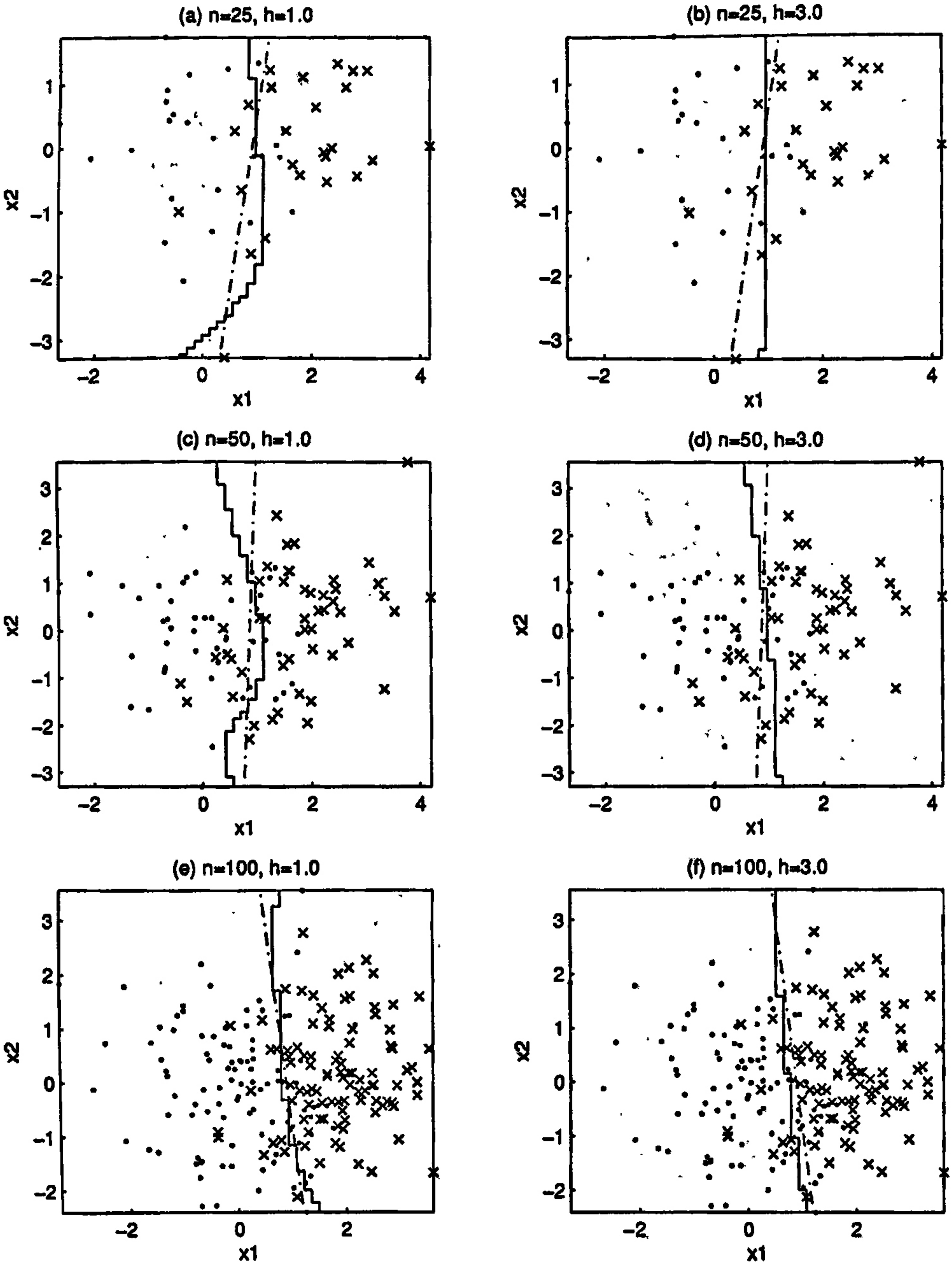


Figure 2.5 : Classification boundaries for fixed bandwidth kernel classifiers and linear classifiers

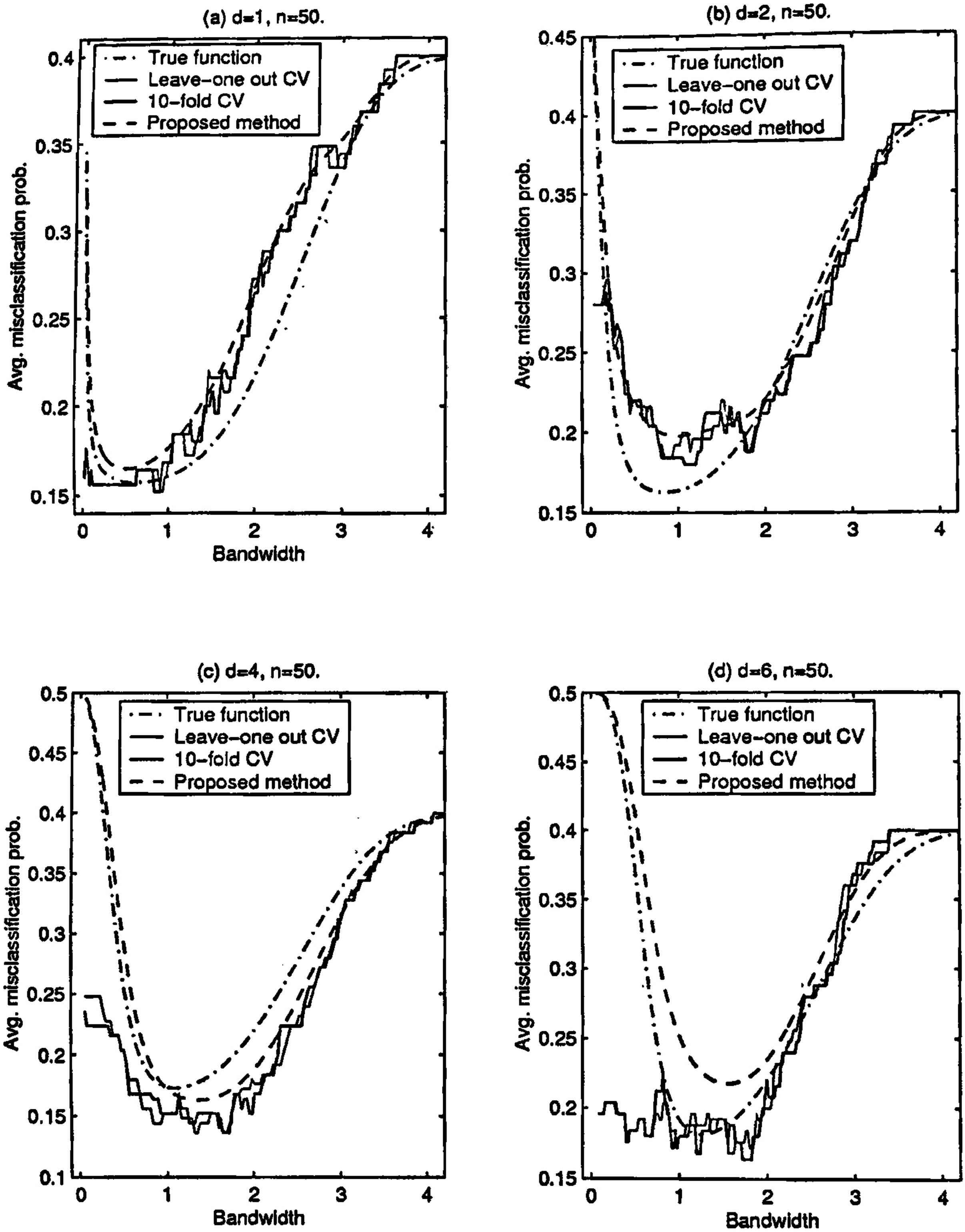


Figure 2.6 : Average misclassification probabilities (unequal prior cases)

our findings reported in the preceding sections. In a J-class discrimination problem, if we use J different bandwidths $h_1, h_2, \ldots, h_J$ for the J populations, average misclassification probability is given by

$$\begin{aligned}\Delta(h_1, h_2, \ldots, h_J) &= \sum_{j=1}^{J} \pi_j \int P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) \leq \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for some } i \neq j\}\ f_j(\mathbf{x})d\mathbf{x} \\ &= 1 - \sum_{j=1}^{J} \pi_j \int P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) > \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for all } i \neq j\}\ f_j(\mathbf{x})d\mathbf{x} \\ &= 1 - \sum_{j=1}^{J} \pi_j \int \left[\int \prod_{i \neq j} P\{\pi_i \hat{f}_{ih_i}(\mathbf{x}) < u\}\ g_{jh_j}(u)\ du \right]\ f_j(\mathbf{x})d\mathbf{x},\end{aligned}$$

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 re-sampling techniques like bootstrap (see e.g., Efron, 1982; 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 (see Section 2.3.1), 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. 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 the 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 $\hat{f}_{jh_j}(\mathbf{x})$ $(j = 1, 2, \ldots, J)$. Then, the average misclassification probability can be approximated by

$$\psi(h_1, h_2, \ldots, h_J) = 1 - \sum_{j=1}^{J} \pi_j \int \left[\int \prod_{i \neq j} \Phi\left\{ \frac{u - \pi_i \mu_{ih_i}(\mathbf{x})}{\pi_i s_{ih_i}(\mathbf{x})} \right\}\ \phi\left\{u, \pi_j \mu_{jh_j}(\mathbf{x}), \pi_j s_{jh_j}(\mathbf{x})\right\}\ du \right]\ f_j(\mathbf{x})d\mathbf{x},$$

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 2.4: *Suppose that $n_1, n_2, \ldots, n_J$ $(N = \sum n_j)$ are the training sample sizes from the J populations, and $h_{n_1}, h_{n_2}, \ldots, h_{n_J}$ are the bandwidths used in kernel estimates of population densities $f_1, f_2, \ldots, f_J$ respectively. Further, assume that the densities $f_1, f_2, \ldots, f_J$ have bounded third derivatives, and the kernel K is bounded and symmetric about 0 satisfying $\int \|\mathbf{y}\|^3 K^2(\mathbf{y})d\mathbf{y} < \infty$. For every $j \in \{1, 2, \ldots, J\}$, as $N \to \infty$, we also assume that*

(i) $h_{n_j} \to 0$ and $h_{n_j}/h_{n_i} \to C_{ji} > 0$ for all i,

(ii) $n_j h_{n_j}^d \to \infty$, *and*
(iii) $n_j/N \to \lambda_j$ *such that* $0 < \lambda_j < 1$.
Then as $N \to \infty$, $\left|\Delta(h_{1_{n_1}}, h_{2_{n_2}}, \ldots, h_{J_{n_J}}) - \psi(h_{1_{n_1}}, h_{2_{n_2}}, \ldots, h_{J_{n_J}})\right| \to 0$, *and both* $\Delta(h_{1_{n_1}}, h_{2_{n_2}}, \ldots, h_{J_{n_J}})$ *and* $\psi(h_{1_{n_1}}, h_{2_{n_2}}, \ldots, h_{J_{n_J}})$ *tend to the optimal Bayes risk.*

Thus, if one minimizes $\psi(h_1, h_2, \ldots, h_J)$ w.r.t. $h_1, h_2, \ldots, h_J$, one has a kernel density estimate based classification rule with asymptotic average misclassification probability equal to the optimal Bayes risk, under suitable regularity conditions.

2.3.1 Data analytic implementation

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

$$\psi_N(h_1, h_2, \ldots, h_J) = 1 - \sum_{j=1}^{J} \frac{\pi_j}{n_j} \sum_{k=1}^{n_j} \left[\int \prod_{i \neq j} \Phi \left\{ \frac{u - \pi_i \mu_{ih_i}(\mathbf{x}_{jk})}{\pi_i s_{ih_i}(\mathbf{x}_{jk})} \right\} \phi \left\{ u, \pi_j \mu_{jh_j}(\mathbf{x}_{jk}), \pi_j s_{jh_j}(\mathbf{x}_{jk}) \right\} du \right],$$

where $\mathbf{x}_{jk}$ is the k^{th} observation of the j^{th} 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; 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 $\psi_N(h_1, h_2, \ldots, h_J)$ by $\psi^*_N(h_1, h_2, \ldots, h_J)$, where

$$\psi^*_N(h_1, h_2, \ldots, h_J) = 1 - \sum_{j=1}^{J} \frac{\pi_j}{n_j} \sum_{k=1}^{n_j} \left[\int \prod_{i \neq j} \Phi \left\{ \frac{u - \pi_i \mu^*_{ih_i}(\mathbf{x}_{jk})}{\pi_i s^*_{ih_i}(\mathbf{x}_{jk})} \right\} \phi \left\{ u, \pi_j \mu^*_{jh_j}(\mathbf{x}_{jk}), \pi_j s^*_{jh_j}(\mathbf{x}_{jk}) \right\} du \right].$$

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

$$\mu^*_{jh_j}(\mathbf{x}_{jk}) = \frac{1}{n_j - 1} \sum_{\substack{l=1 \\ l \neq k}}^{n_j} \phi_d \left(\mathbf{x}_{jk}, \mathbf{x}_{jl}, \{h_j^2 + h_{oj}^2\} \mathrm{I}_d \right) \quad \text{and}$$

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

where $\phi_d(\mathbf{x}, \mu, \mathbf{\Sigma}) = (2\pi)^{-d/2}|\mathbf{\Sigma}|^{-1/2}\exp\{-(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)/2\}$, and h_{oj} is the bandwidth that minimizes estimated $MISE$ of a kernel density estimate of the j^{th} 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 $1 \leq l \leq n_i$ (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, $\psi^*_N(h_1, h_2, \ldots, h_J)$ is a nice smooth function, and we propose to minimize it over $h_1, h_2, \ldots, h_J$ to find out the optimal bandwidths. Note that, unlike the step function obtained in V-fold cross-validation, $\psi^*_N(h_1, h_2, \ldots, h_J)$ is a smooth function, and we propose to minimize it over $h_1, h_2, \ldots, h_J$ to find optimal bandwidths. In all the examples considered in Figures 2.2 and 2.6, 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.

2.4 Results from simulation experiments

In this section, we report on 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 μ_1 and μ_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_o) 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 the Tables 2.1-2.4. 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 2.1. 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 method.

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 2.1 : Misclassification rates : normal 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 2.2 : Misclassification rates : double exponential distributions with equal 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 2.3 : Misclassification rates : normal distributions with unequal priors

Table 2.2 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_o, but had slightly higher error rates and substantially worse standard errors than our proposed method.

Bandwidth selection is more critical when the priors are not equal. To evaluate the performance of the proposed method in such situations, we considered the same examples (as above) but set π_1=0.6 and 0.7 respectively. For μ =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 2.3 & 2.4). These results again show good performance of our method as a bandwidth selector for both normal and double exponential populations.

π_1	Bayes risk	d	n	Δ in percentage				
				h_o	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)

Table 2.4 : Misclassification rates : double exponential distributions with unequal priors

2.5 Results from the analysis of benchmark data sets

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 1000 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 rates over these 1000 random splits are 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) in Figures 2.7 and 2.8. 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.

We begin with Fisher's (1936) Iris data, where four measurements (sepal length, sepal width, petal length and petal width) are taken on each observation coming from one of the three populations of Iris plants : 'Iris Setosa', 'Iris Virginica' and 'Iris Versicolor'. This data set was collected by Dr. Edgar Anderson and Fisher used it to demonstrate the utility of the discriminant functions. It is available from http://lib.stat.cmu.edu. 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 for different populations. Of course, it is possible to use other methods of standardization (see e.g., Coolie and MacEachern, 1998).

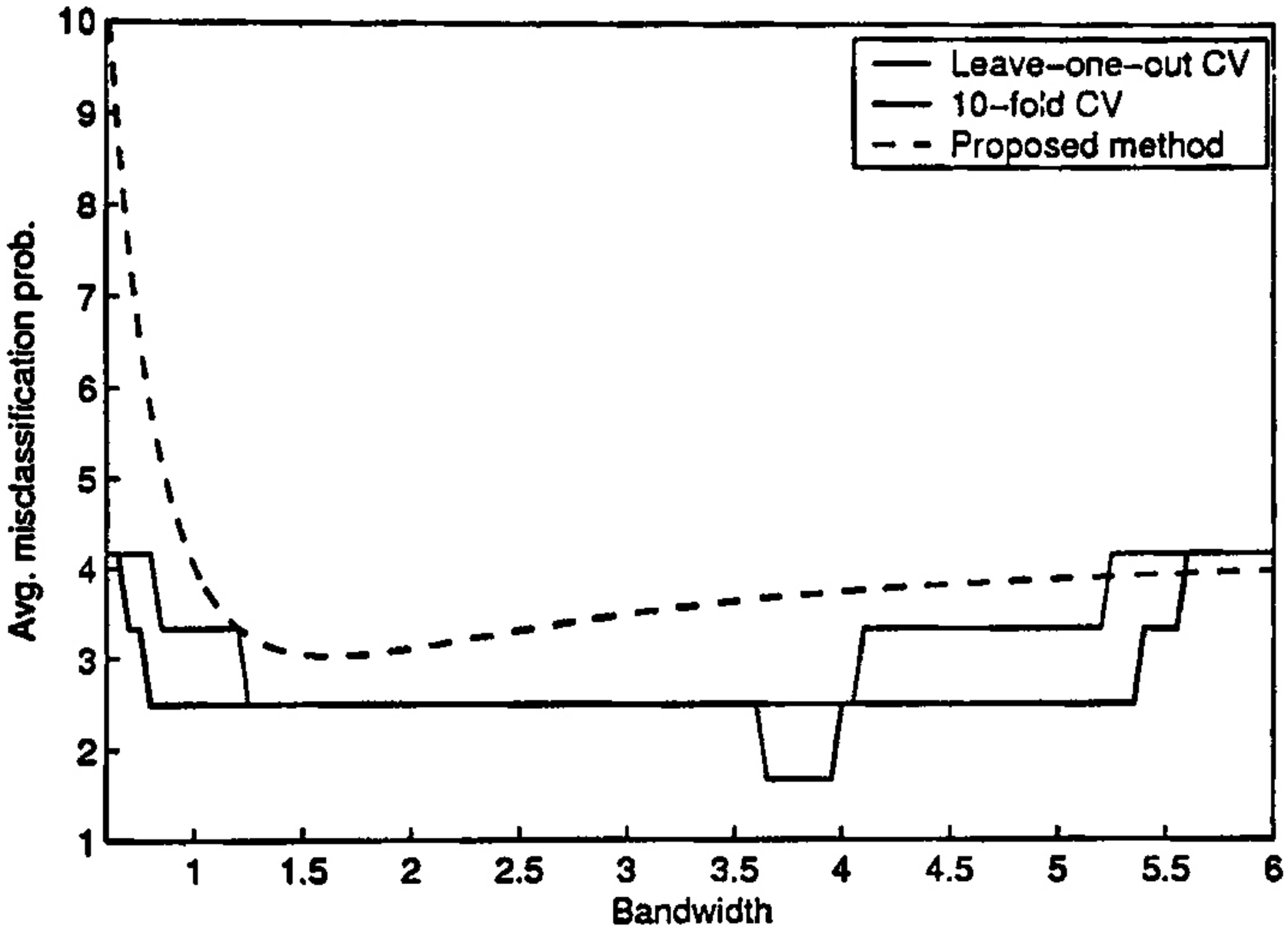


Figure 2.7 : Average misclassification probabilities (Iris data)

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%. When bandwidths for different population densities were estimated by the usual leave-one-out least squares cross-validation technique (see e.g., Silverman, 1986; 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 higher than that for simple linear discriminant analysis. However, for both cross-validation methods (leave-one-out and 10-fold) 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% with a standard error of 0.09%.

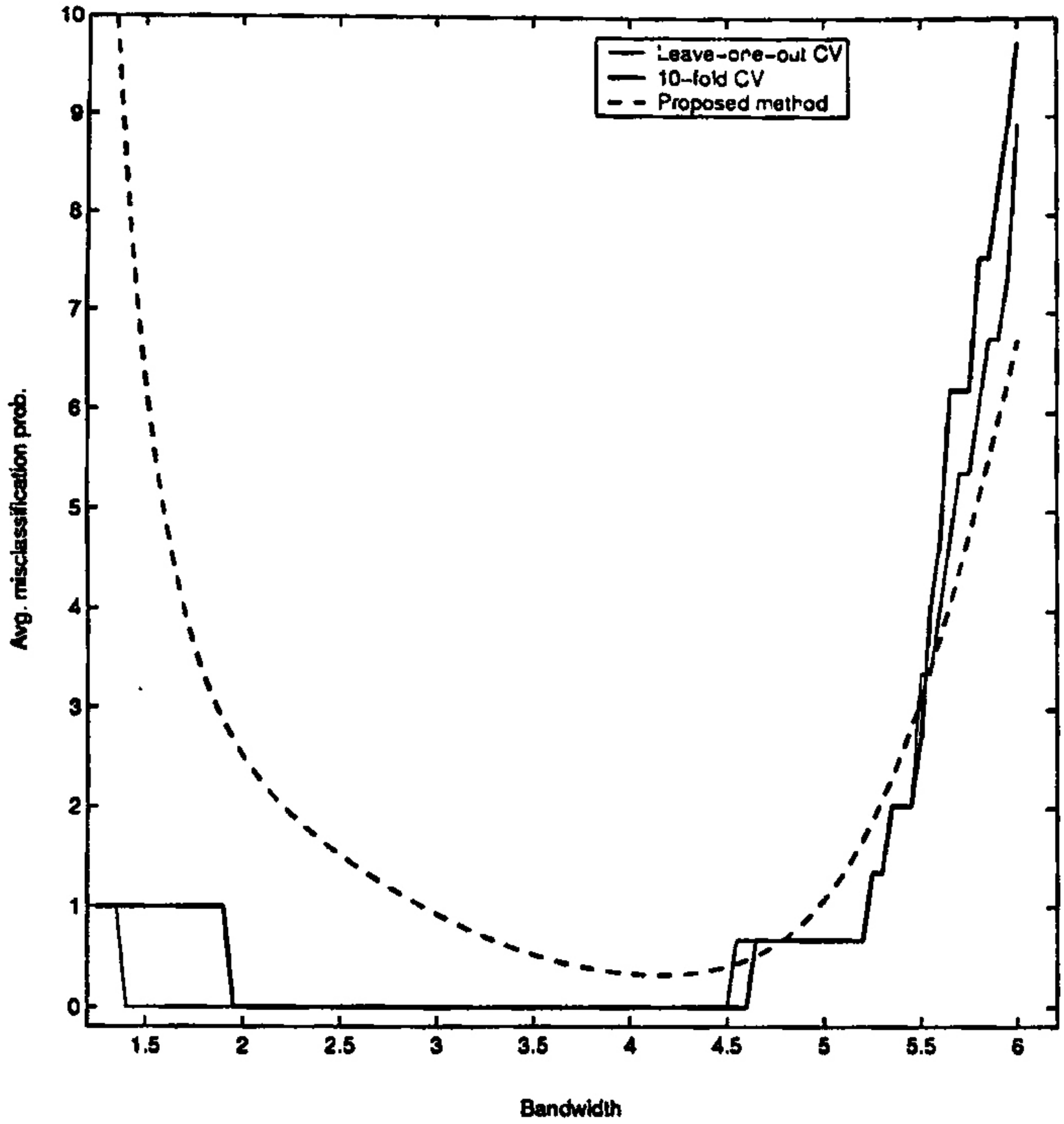


Figure 2.8 : Average misclassification probabilities (wine data)

The other data set that we have analyzed is known as 'wine data'. This data set is available from http://www.uci.ics.edu. It contains the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivators. 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. It 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 data set were used to estimate these prior probabilities. Traditional 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 bought the error rate to 0.48% with a standard error of 0.04%.

2.6 Remarks and discussions

In this chapter, we have pointed out some fundamental differences between the optimal bandwidth that minimizes the $MISE$ of a kernel density estimate and the optimal bandwidth that minimizes the average misclassification probability when kernel density estimates are used for discriminant analysis. One of our main findings is that depending on whether the prior probabilities for different populations are equal or not, average misclassification probability exhibits totally different behaviors for varying choices of the bandwidth parameter. While it is a common belief that with the growing sample size, one should gradually shrink the bandwidth to zero to achieve better performance of the kernel density estimate, we have shown that when the prior probabilities are all equal, depending on nature of the population density functions and the kernel, large bandwidths may also lead to nearly optimal misclassification rates.

In the equal prior cases, we have further shown that for large bandwidths, the average misclassification rate for a kernel density estimate based classifier asymptotically tends to that for a linear discrimination rule, which is optimal under certain assumptions on the underlying population densities. This result is analogous to what is known for regression problems where a nonparametric method (e.g. local polynomial regression) tends to a parametric regression technique (e.g. usual polynomial regression) as the bandwidth gets larger.

Our analysis of one of the most well known data, namely Fisher's Iris data, nicely demonstrates the importance of proper choice of the bandwidth parameter for kernel density estimate based discriminant analysis. As we have seen in the preceding section, a choice of the bandwidth that is motivated by minimization of the $MISE$ may lead to a performance of the classifier which is worse than that of simple linear classifier. However, a more judicious choice of the bandwidth will enable the kernel density estimate based classifier to outperform the linear classifier.

We have adequately demonstrated that V-fold cross-validation type techniques often do not lead to a unique optimal choice of the bandwidth primarily due to step function like shape of the estimated error curve with multiple minima. However, in some cases, the largest minimizer of the estimated misclassification probability may lead to a decent performance. As we have already discussed, since the kernel density estimate is an average of i.i.d. random variables, normal approximation to its distribution works extremely well with very high degree of accuracy for large or even moderately large sample sizes. Our proposed

method of bandwidth selection has essentially tried to exploit this by using normal approximations to certain probabilities instead of crude empirical proportions. This has resulted in smooth estimates of error curves, which typically have unique minima and resemble the true average misclassification probability curves better than the V-fold cross-validation curves in simulated examples.

2.7 Proofs and mathematical details

Proof of Theorem 2.1 : For all $j \in \{1, 2, \ldots, J\}$, $\hat{f}_{jh}(\mathbf{x}) = n^{-1}h^{-d}\sum_{k=1}^{n_j} K\{(\mathbf{x} - \mathbf{x}_{jk})/h\}$ 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 the kernel density estimates will be same as that of the classification rule based on the theoretical scale space functions $E\{\hat{f}_{1h}(\mathbf{x})\}$, $E\{\hat{f}_{2h}(\mathbf{x})\}$,..., $E\{\hat{f}_{Jh}(\mathbf{x})\}$ associated with population densities $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_J(\mathbf{x})$ 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 μ_j as the center of symmetry.

Now, choose $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\|\mathbf{x}_1 - \mu_j\| < \|\mathbf{x}_2 - \mu_j\|$ (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 whole d-dimensional space into two half-spaces (H^+ and H^-). It is clear that $\mathbf{x}_1$ and μ_j belong to the same half-space (let us denote it by H^+) and $\mathbf{x}_2$ to the other (call it H^-). 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

$$K\{(\mathbf{x}_1 - \mathbf{y}^*)/h\} - K\{(\mathbf{x}_2 - \mathbf{y}^*)/h\} = K\{(\mathbf{x}_2 - \mathbf{y})/h\} - K\{(\mathbf{x}_1 - \mathbf{y})/h\} \quad > 0.$$

Therefore, whatever may be the value of h,

$$h^{-d} \int_{\mathbf{y}^* \in H^+} f_j(\mathbf{y}^*)\,[K\{(\mathbf{x}_1 - \mathbf{y}^*)/h\} - K\{(\mathbf{x}_2 - \mathbf{y}^*)/h\}]\,d\mathbf{y}^*$$

$$> \; h^{-d} \int_{\mathbf{y} \in H^-} f_j(\mathbf{y})\,[K\{(\mathbf{x}_2 - \mathbf{y})/h\} - K\{(\mathbf{x}_1 - \mathbf{y})/h\}]\,d\mathbf{y}.$$

$$\Rightarrow h^{-d} \int_{\mathbf{y} \in H^+} f_j(\mathbf{y})\,[K\{(\mathbf{x}_1 - \mathbf{y})/h\} - K\{(\mathbf{x}_2 - \mathbf{y})/h\}]\,d\mathbf{y}$$

$$+ \; h^{-d} \int_{\mathbf{y} \in H^-} f_j(\mathbf{y})\,[K\{(\mathbf{x}_1 - \mathbf{y})/h\} - K\{(\mathbf{x}_2 - \mathbf{y})/h\}]\,d\mathbf{y} > 0.$$

$$\Rightarrow E\{\hat{f}_{j_h}(\mathbf{x}_1)\}-E\{\hat{f}_{j_h}(\mathbf{x}_2)\} \;=\; h^{-d}\int\limits_{\mathbf{y}\in R^d} f_j(\mathbf{y})\left[K\{(\mathbf{x}_1-\mathbf{y})/h\}-K\{(\mathbf{x}_2-\mathbf{y})/h\}\right]d\mathbf{y} > 0.$$

So, the convolution is also a decreasing function of the distance from its center of symmetry. Now, $f_i(\mathbf{x}) > f_j(\mathbf{x}) \Leftrightarrow \|(\mathbf{x}-\mu_i)\|^2 < \|(\mathbf{x}-\mu_j)\|^2 \Leftrightarrow E\{\hat{f}_{ih}(\mathbf{x})\} > E\{\hat{f}_{jh}(\mathbf{x})\}$ since the distributions satisfy the location shift model. Hence, for all $h > 0$, the 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 0 and bounded third derivatives. Then, as $h \to \infty$, the expectation and the variance of $\hat{f}_h(\mathbf{x})$ are given by

$$\begin{aligned} E\{\hat{f}_h(\mathbf{x})\} &= h^{-d}[K(0)+(1/2h^2)E_f\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\}+O(h^{-3})] \text{ and} \\ Var\{\hat{f}_h(\mathbf{x})\} &= (4nh^{2d+4})^{-1}[Var_f\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\}+O(h^{-1})]. \end{aligned}$$

Proof of Proposition 2.1 : The expectation and the variance of $\hat{f}_h(\mathbf{x})$ can be written as

$$E\{\hat{f}_h(\mathbf{x})\} = h^{-d}E_f\left[K\{(\mathbf{x}-\mathbf{X})/h\}\right] \text{ and } Var\{\hat{f}_h(\mathbf{x})\} = n^{-1}h^{-2d}Var_f\left[K\{(\mathbf{x}-\mathbf{X})/h\}\right].$$

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

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

where $Y_{i,j,k} = (x_i-X_i)(x_j-X_j)(x_k-X_k)\frac{\partial^3 K(\mathbf{t})}{\partial t_i \partial t_j \partial t_k}|_{\mathbf{t}=\boldsymbol{\xi}}$ for some intermediate vector $\boldsymbol{\xi}$ between 0 and $(\mathbf{x}-\mathbf{X})/h$. Therefore, using the fact that K has bounded third derivatives and $\int \|\mathbf{x}\|^6 f(\mathbf{x})d\mathbf{x} < \infty$, we get

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

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

(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 : For the ease of notation, 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 $s_h^2(\mathbf{x}) = Var\{Y_h(\mathbf{x})\}$.

(a) When $\pi_1 = \pi_2$, it is evident from Proposition 2.1 that

(i) as $h \to \infty$, the sign of $\mu_h(\mathbf{x})$ and the sign of
$\mathbf{x}'\nabla^2 K(0)\{E_{f_2}(\mathbf{X}) - E_{f_1}(\mathbf{X})\} - (1/2)\left[E_{f_2}\left\{\mathbf{X}'\nabla^2 K(0)\mathbf{X}\right\} - E_{f_1}\left\{\mathbf{X}'\nabla^2 K(0)\mathbf{X}\right\}\right]$
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, from Chebychev's inequality,

$$\begin{aligned} P\{Y_h(\mathbf{x}) \leq 0\} &\geq \frac{\mu_h^2(\mathbf{x})}{\mu_h^2(\mathbf{x}) + s_h^2(\mathbf{x})} \quad \text{when } \mu_h(\mathbf{x}) \leq 0 \\ &\leq \frac{s_h^2(\mathbf{x})}{\mu_h^2(\mathbf{x}) + s_h^2(\mathbf{x})} \quad \text{when } \mu_h(\mathbf{x}) > 0. \end{aligned}$$

As $n, h \to \infty$, the right side tends to 1 and 0, respectively, in the first and the second cases. Therefore, $P\{Y_h(\mathbf{x}) \leq 0\}$ also tends to 1 and 0 in the respective cases.

(b) 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 Chebychev's inequality, $\lim_{h \to \infty} P\{Y_h(\mathbf{x}) \leq 0\} = 0$.

Proof of Theorem 2.2 : For $1 \leq i \neq j \leq J$, 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 $\sum_{j:\, j \neq i} P(A_{ij}^h) - (J-2) \leq P\left\{\bigcap_{j:\, j \neq i} A_{ij}^h\right\} \leq \min_{j \neq i} P(A_{ij}^h)$. Now, it is easy to see that, for any i, $P\left\{\bigcap_{j:\, j \neq i} A_{ij}^h\right\} \to 1$ if and only if $P(A_{ij}^h) \to 1$ for all $j \neq i$. The proof of the theorem then follows from part (a) of Lemma 2.1.

Proof of Theorem 2.3 : Since the population densities satisfy the location shift model (with location parameter μ_j for the j^{th} class), for $i \neq j$, $E_{f_i}(\mathbf{X}) - E_{f_j}(\mathbf{X}) = \mu_i - \mu_j$ and $E_{f_i}\{\mathbf{X}'\nabla^2 K(0)\mathbf{X}\} - E_{f_j}\{\mathbf{X}'\nabla^2 K(0)\mathbf{X}\} = \mu_i'\nabla^2 K(0)\mu_i - \mu_j'\nabla^2 K(0)\mu_j$. The proof of (a) is now immediate from Theorem 2.2. The proofs of (b) and (c) follow from Lemma 2.1, using the same logic as in the proof of Theorem 2.2.

Lemma 2.2: Assume that the density function f has bounded third derivatives and the kernel K is symmetric about 0 with $\int \|\mathbf{y}\|^3 K^2(\mathbf{y})d\mathbf{y} < \infty$. Then, as $h \to 0$, the expectation and the variance of $\hat{f}_h(\mathbf{x})$ is given by

$$E\{\hat{f}_h(\mathbf{x})\} = f(\mathbf{x}) + O(h^2), \quad Var\{\hat{f}_h(\mathbf{x})\} = (nh^d)^{-1}\{\beta f(\mathbf{x}) + O(h^2)\}, \text{ where } \beta = \int K^2(\mathbf{y})d\mathbf{y}.$$

Proof of Lemma 2.2 :

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

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, $E_f\left[h^{-2d}K^2\{(\mathbf{x}-\mathbf{X})/h)\}\right] = h^{-d}\int K^2(\mathbf{y})\ [f(\mathbf{x}) - h\{\mathbf{y}'\nabla f(\mathbf{x})\}$

$$+ (h^2/2)\{\mathbf{y}'\nabla^2 f(\mathbf{x})\mathbf{y}\} + (h^3/3)\sum_{i,j,k} Z_{i,j,k}]\ d\mathbf{y}$$

$$= h^{-d}[\beta f(\mathbf{x}) + O(h^2)], \quad \text{where } \beta = \int K^2(\mathbf{y})d\mathbf{y}.$$

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

Lemma 2.3 : Suppose that $n_1, n_2, \ldots, n_J$ $(N = \sum n_j)$ are the training sample sizes from the J populations and $h_{n_1}, h_{n_2}, \ldots, h_{n_J}$ are the bandwidths used in kernel estimates of population densities $f_1, f_2, \ldots, f_j$ respectively. Further, assume that the densities $f_1, f_2, \ldots, f_j$ and the kernel K satisfy the conditions assumed in Lemma 2.2. For every $j \in \{1, 2, \ldots, J\}$ and $N \to \infty$, we also assume that

(*i*) $h_{n_j} \to 0$ and $h_{n_j}/h_{n_i} \to C_{ji} > 0$ for all i,

(*ii*) $n_j h_{n_j}^d \to \infty$ and

(*iii*) $n_j/N \to \lambda_j$ such that $0 < \lambda_j < 1$.

For some $\mathbf{x} \in R^d$, define $X_1, X_2, \ldots, X_J$ as independently distributed normal variates with $E(X_i) = \pi_i\mu_{ih_{n_i}} = E\{\pi_i\hat{f}_{ih_{n_i}}(\mathbf{x})\}$ and $Var(X_i) = \pi_i^2 s_{ih_{n_i}}^2 = Var\{\pi_i\hat{f}_{ih_{n_i}}(\mathbf{x})\}$.

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

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

Proof of Lemma 2.3 :

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

Let $\theta_N(x) = \prod_{i \neq 1} \Phi\left(\frac{x - \pi_i \mu_{ih_{n_i}}}{\pi_i s_{ih_{n_i}}}\right)$. Using a Taylor expansion about $\pi_1 \mu_{1h_{n_1}}$, $\theta_N(x)$ can be expressed as $\theta_N(x) = \theta_N(\pi_1 \mu_{1h_{n_1}}) + (x - \pi_1 \mu_{1h_{n_1}})\ \theta'_N(\xi)$ for some value ξ that lies between $\pi_1 \mu_{1h_{n_1}}$ and x.

We have $(x - \pi_1 \mu_{1h_{n_1}})\ \theta'_N(\xi) = \sum_{j=2}^{J} \frac{x - \pi_1 \mu_{1h_{n_1}}}{\pi_j s_{jh_{n_j}}}\ \beta_{jh_{n_j}}(\xi)$, where

$$\beta_{jh_{n_j}}(\xi) = \phi\left(\frac{\xi - \pi_j \mu_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right) \prod_{k:\ k \neq 1, j} \Phi\left(\frac{\xi - \pi_k \mu_{kh_{n_k}}}{\pi_k s_{kh_{n_k}}}\right).$$

Now, $E\left(\left|\frac{x - \pi_1 \mu_{1h_{n_1}}}{\pi_j s_{jh_{n_1}}}\ \beta_{jh_{n_j}}(\xi)\right|\right) \leq E\left\{\left|\frac{x - \pi_1 \mu_{1h_{n_1}}}{\pi_j s_{jh_{n_j}}}\ \phi\left(\frac{\xi - \pi_j \mu_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right)\right|\right\}$ (since $|\Phi(\cdot)| \leq 1$)

$$\leq E^{1/2}\left(\frac{x - \pi_1 \mu_{1h_{n_1}}}{\pi_j s_{jh_{n_j}}}\right)^2 E^{1/2}\left\{\phi^2\left(\frac{\xi - \pi_j \mu_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right)\right\}$$
$$= (\pi_1^2/\pi_j^2)(s^2_{1h_{n_1}}/s^2_{jh_{n_j}}) E^{1/2}\left\{\phi^2\left(\frac{\xi - \pi_j \mu_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right)\right\}.$$

As $N \to \infty$, under the given conditions, both $s^2_{1h_{n_1}}$ and $s^2_{jh_{n_j}}$ tend to 0 (from Lemma 2.2) but $s^2_{1h_{n_1}}/s^2_{jh_{n_j}}$ tends to a positive constant. Therefore, as $N \to \infty$, $\phi^2\left(\frac{\xi - \pi_j \mu_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right) \to 0$
$\Rightarrow E\left\{\phi^2\left(\frac{\xi - \pi_j \mu_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right)\right\} \to 0$ (by Dominated Convergence Theorem)
$\Rightarrow E|(x - \pi_1 \mu_{1h_{n_1}})\ \theta'_N(\xi)| \to 0.$

Proof of Theorem 2.4 : For a J-class problem, $\psi(h_{n_1}, h_{n_2}, \ldots, h_{n_J})$ is given by

$\psi(h_{n_1}, h_{n_2}, \ldots, h_{n_J})$

$$= 1 - \sum_{j=1}^{J} \pi_j \int \left[\int \prod_{i \neq j} \Phi\left\{\frac{u - \pi_i \mu_{ih_{n_i}}(\mathbf{x})}{\pi_i s_{ih_{n_i}}(\mathbf{x})}\right\}\ \phi\left\{u, \pi_j \mu_{jh_{n_j}}(\mathbf{x}), \pi_j s_{jh_{n_j}}(\mathbf{x})\right\}\ du\right]\ f_j(\mathbf{x})\ d\mathbf{x},$$

where $\pi_j \mu_{jh_{n_j}}(\mathbf{x})$ 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 2.3 implies that, as $N \to \infty$, for all j,

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

Also, from Lemma 2.2, for every j, as $N \to \infty$,

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

$$\Rightarrow \psi(h_{n_1}, h_{n_2}, \ldots, h_{n_J}) \to 1 - \sum_{j=1}^{J} \pi_j \int_{f_j > f_i \ \forall i \neq j} f_j(\mathbf{x}) d\mathbf{x}, \quad \text{(by Dominated Convergence Theorem), which is the optimal Bayes risk.}$$

In view of the asymptotic orders of $E\{\hat{f}_{jh_{n_j}}(\mathbf{x})\}$ and $Var\{\hat{f}_{jh_{n_j}}(\mathbf{x})\}$ obtained in Lemma 2.2, Lindeberg's condition for Multivariate Central Limit Theorem holds for

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

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

$$\left| P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) > \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for all } i \neq j\} \right.$$

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

as $N \to \infty$ using the results in Bhattacharya and Ranga Rao (1976, Section 2, pp. 6-23) on uniform convergence to multivariate normal probabilities for convex sets with boundaries having zero Lebesgue measure. Finally, by dominated convergence theorem, we have

$$|\Delta(h_{n_1}, h_{n_2}, \ldots, h_{n_J}) - \psi(h_{n_1}, h_{n_2}, \ldots, h_{n_J})| \to 0 \text{ as } N \to \infty.$$

Chapter 3

Multi-scale kernel discriminant analysis and visualization

3.1 Main problem and motivation

Classification based on kernel density estimates has been widely discussed in the existing literature on pattern recognition and statistical learning (see, e.g. Devijver and Kittler, 1982; Devroye, Gyorfi and Lugosi, 1996; Duda *et. al.*, 2000; Hastie *et. al.*, 2001). In these classification methods, one uses the kernel estimates of the unknown population densities $f_j(\mathbf{x})$ $(j = 1, 2, \ldots, J)$ and plug them in the Bayes rule to build the classifier of the form $\mathrm{d}_K(\mathbf{x}) = \arg\max_j \pi_j \hat{f}_{jh_j}(\mathbf{x})$, where π_j's are the prior probabilities and $\hat{f}_{jh_j}(\mathbf{x})$'s are the kernel density estimates of the respective classes. Recall that if $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are d-dimensional observations in the training sample from the j^{th} population $(j = 1, 2, \ldots, J)$, the kernel density estimate $\hat{f}_{jh_j}(\mathbf{x})$ of the j^{th} population is given by $\hat{f}_{jh_j}(\mathbf{x}) = n_j^{-1} h_j^{-d} \sum_{k=1}^{n_j} K\left\{h_j^{-1}(\mathbf{x}_{jk} - \mathbf{x})\right\}$, where $K(\cdot)$ is a d-dimensional density function, and $h_j > 0$ is the bandwidth parameter.

As we have observed earlier (in Chapter 2), the performance of a classifier based on kernel density estimates depends critically on the values of bandwidth parameters. Existing data based bandwidth selection techniques (see e.g., Silverman, 1986; Hall and Marron, 1987; Sheather and Jones, 1991; Jones, Marron and Sheather, 1996b) that target to minimize the mean integrated square error ($MISE$) for the density estimate, are not good for discriminant analysis problems, and they may lead to rather poor misclassification rates for the resulting classifier (see Section 2.1 and 2.2). On the other hand, popular V-fold cross-validation (see e.g., Stone [M. Stone], 1977) and similar methods for bandwidth selection in a nonparametric classification problem are not quite effective due to piecewise constant nature of estimated misclassification probability functions with infinitely many minima.

Further, all such cross-validation based techniques require a huge computation when there are several competing classes. Two other important points to keep in mind in the case of discriminant analysis using kernel density estimates are the following :

(i) The choice of bandwidths should depend on the specific observation to be classified in addition to depending on the population densities, and given a specific observation to be classified, one needs to assess the strength of the evidence in favor of one population or the other for varying choices of bandwidths for density estimates corresponding to different competing populations.

(ii) In a multi-class discrimination problem, instead of using a single bandwidth for each population density estimate, it is more meaningful to use different bandwidths for a class density estimate when it is compared with the density estimates for different competing classes for classifying a specific observation.

Instead of choosing a single optimal bandwidth for each class density estimate, in this chapter we consider a family of density estimates $\{\hat{f}_{jh_j} : h_j \in \mathcal{H}_j\}$ for each population over a wide range of bandwidths to carry out a multi-scale version of kernel discriminant analysis (see also, Chaudhuri and Marron, 1999, 2000; Godtliebsen, Marron and Chaudhuri, 2002). Simultaneous consideration of different levels of the smoothing is expected to yield more information useful for classification purpose than that obtained in an approach based on a single optimum bandwidth for each class density estimate. The results of the analysis are presented using two-dimensional plots, which are specific to an observation to be classified, and there one can visually compare the strength of the evidence in favor of different competing classes over wide ranges of smoothing parameters. Statistical uncertainties at various locations in the plots are also quantified on the basis of appropriately estimated misclassification probabilities, and they too are displayed using some two-dimensional plots to facilitate the decision about classification. Of course, the final decision about classifying an observation is to be made by some judicious combination of all the information obtained at different levels of smoothing, and we will discuss some appropriate ways for doing that.

3.2 Description of multi-scale methodology

Suppose that $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are the training sample observations from the j^{th} class, where $1 \leq j \leq J$. In order to classify an observation $\mathbf{x}$ into one of the J classes, we first need to obtain the density estimates $\hat{f}_{jh_j}(\mathbf{x})$ at the point $\mathbf{x}$ for all $j = 1, 2, \ldots, J$. In practice, before computing the class density estimate, we can standardize the data points in a class using an estimate of the class dispersion matrix to make the data more spherical in nature and thereby making the use of a common bandwidth h_j for all co-ordinate variables more

justified. The density estimate for the original data vectors can be obtained from that of the standardized data vectors by using the simple transformation formula for a probability density function when the random vectors undergo a linear transformation. Then, for a given pair of competing classes, say, class-1 and class-2, and a fixed pair of bandwidths h_1 and h_2 for the two class density estimates, there is an ordering between the functions $\pi_1 \hat{f}_{1h_1}(\mathbf{x})$ and $\pi_2 \hat{f}_{2h_2}(\mathbf{x})$ that determines which one of the two classes is more favorable. We now introduce some measures for the strength of this evidence in favor of one class or the other.

3.2.1 Posterior probability

In a two-class problem, for a given observation $\mathbf{x}$ and a given pair of bandwidths h_1 and h_2, the posterior probability in favor of the first population can be given as

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

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

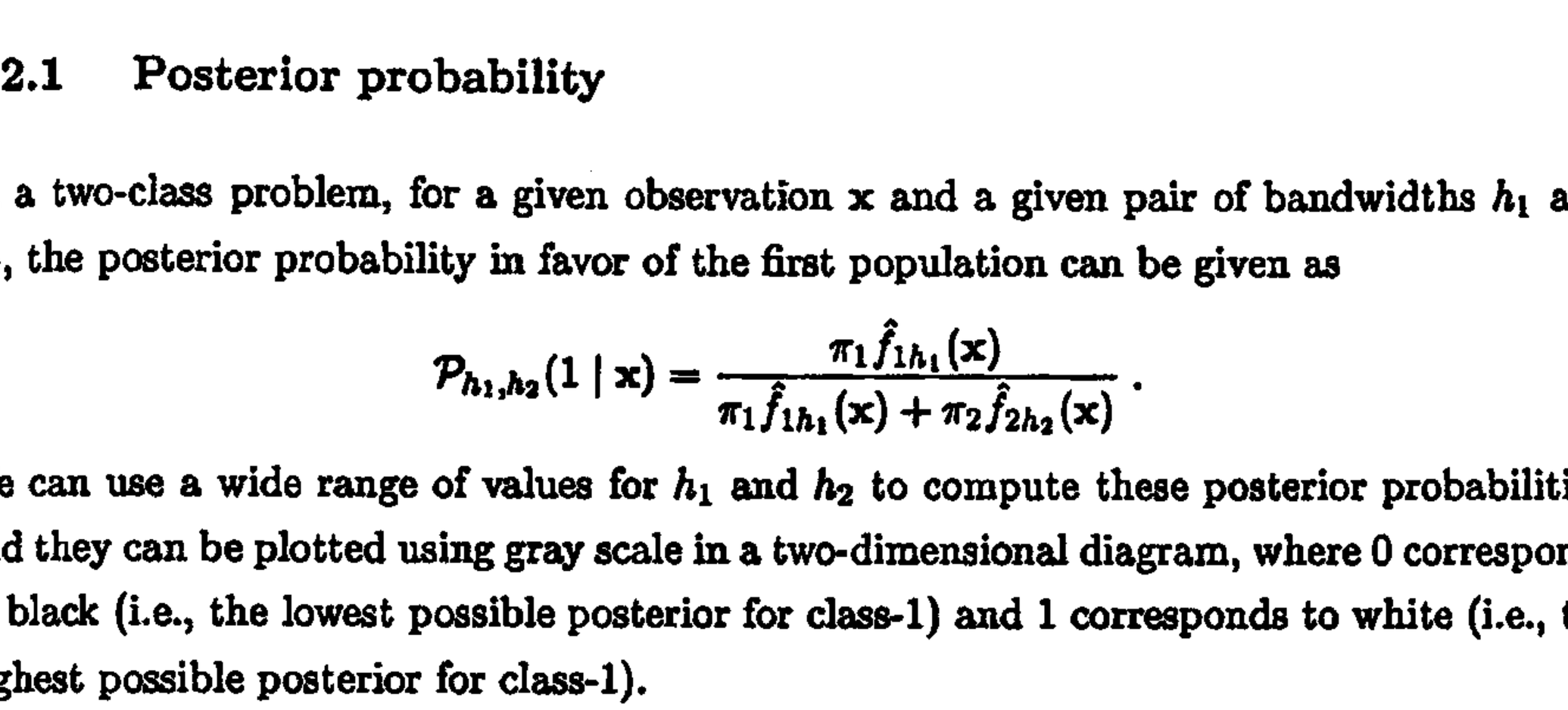


Figure 3.1 : Scatter plots for synthetic data

To demonstrate our methodology, we consider an example data set from Ripley (1994), which is popularly known as the "synthetic data". This bivariate data set is related to a two-class problem, where both the classes are equal mixtures of two bivariate normal distributions. All these normal distributions have the same dispersion structure, and they differ only in their location parameters. This data set contains a training sample of size 250 (125 for each class) and a test set consisting of 1000 observations (500 from each class).

It is available from http://lib.stat.cmu.edu. Scatter plots for the training and the test set are given in Figure 3.1, where the dots (·) and the crosses (×) represent the observations coming form the two classes.

We have chosen three observations (indicated by 'o' in Figure 3.1) from the test set and labelled them as 'A', 'B' and 'C'. These three points are purposively chosen from three different parts of the data. Observation 'A' lies well within the cluster of the observations from population-1, whereas 'C' clearly belongs to population-2. The observation 'B' is taken near the class boundary where both the populations have more or less equal strength. We performed traditional linear (LDA) and quadratic discriminant analysis (QDA) on this data to classify the entire test set observations using the training sample. There are some observations that got misclassified by both the methods, and 'B' is one of them. Though it is originally from population-1, both LDA and QDA gave decisions in favor of population-2. Of course, both of 'A' and 'C' were correctly classified by the linear and the quadratic classifiers. Here, we have used a wide range of bandwidths for each of the two populations to evaluate the posterior.probabilities $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ for different levels of smoothing. As there are equal number of observations for each of the classes, the prior probabilities for our analysis are taken to be equal.

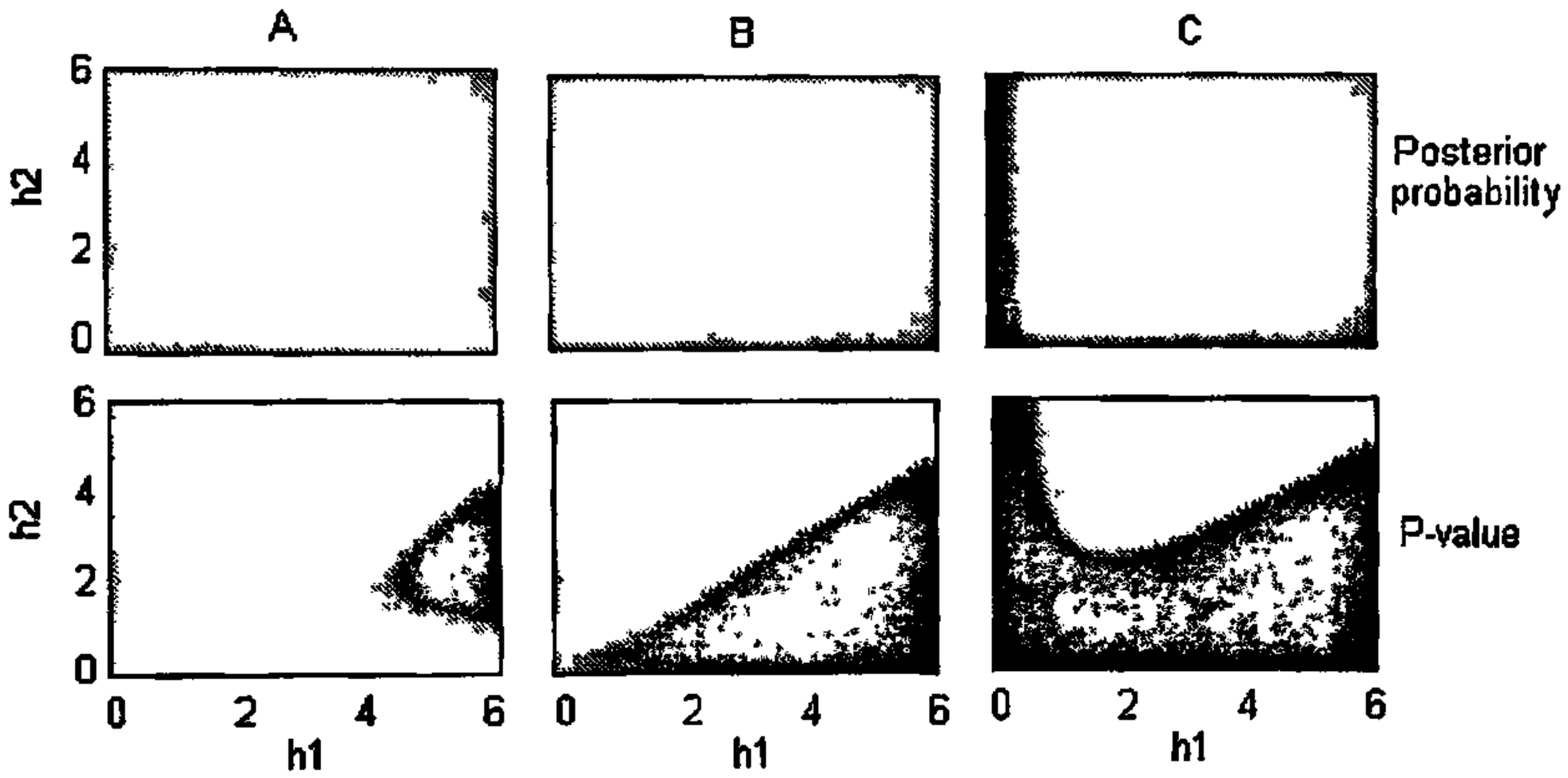


Figure 3.2 : Multi-scale analysis of synthetic data

The top row of Figure 3.2 gives the grey scale representation of posterior probabilities for these three cases, where the bandwidths of the first and the second populations are plotted along the horizontal and the vertical axes, respectively. Though we have allowed h_1 and h_2 to vary in the range $(0, 6)$, one may use longer intervals as well. Since most of the standardized observations are expected to lie within an interval of 6 units (mean $\pm$ 3), it is a fairly good choice for the upper limit of bandwidths. Our empirical experience suggests that further extension of this interval only increases the computational burden but does

not reveal any new pattern. Here white color (high posterior) indicates the regions in favor of the first population whereas the black color (low posterior) points towards the other. Intensity of the color varies with the magnitudes of the posterior probabilities, and this helps us to find out the regions for strong evidence in favor of one of the two populations. As it is expected, we observe a dominance of light colored region in the case of observation 'A' and that of the dark region in the case of observation 'C'. However, for observation 'B', which lies near the class boundary, the evidence is not so clear in favor of any of the two populations. In all our numerical work in this chapter we have used the Gaussian kernel $K(t) = (2\pi)^{-1/2} e^{-t^2/2}$.

3.2.2 A p-value type discrimination measure

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

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

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

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

where Φ is the standard normal distribution function, n_1 and n_2 are the training sample sizes for the two classes, and $\hat{s}^2_{jh_j}(\mathbf{x})$ is the estimated variance of $\hat{f}_{jh_j}(\mathbf{x})$ $(j = 1, 2)$ obtained from the training sample using the sample variance of $h_j^{-d} K\{h_j^{-1}(\mathbf{x}_{j1} - \mathbf{x})\}, h_j^{-d} K\{h_j^{-1}(\mathbf{x}_{j2} - \mathbf{x})\}, \ldots, h_j^{-d} K\{h_j^{-1}(\mathbf{x}_{jn_j} - \mathbf{x})\}$.

An alternative interesting interpretation of the above normal approximation of $P_{h_1,h_2}(\mathbf{x})$ can be given as follows. For a given observation $\mathbf{x}$ and a pair of bandwidths h_1 and h_2, let us imagine a pair of hypotheses $H_0 : \pi_1 E\{\hat{f}_{1h}(\mathbf{x})\} \geq \pi_2 E\{\hat{f}_{2h}(\mathbf{x})\}$ and $H_A : \pi_1 E\{\hat{f}_{1h}(\mathbf{x})\} < \pi_2 E\{\hat{f}_{2h}(\mathbf{x})\}$. If the training sample is used to test these hypotheses using kernel density estimates, which can be viewed as statistics like sample means used in two-sample problems, then the above normal approximation can be taken as the one-sided

p-value associated with that testing problem. This is why we have chosen to call it a p-value type measure of the strength of discrimination.

In the bottom row in Figure 3.2, we have plotted these p-values in two-dimensional plots for the observations 'A' , 'B' and 'C' using gray scales for various choices of values for the bandwidths h_1 and h_2. Like before, the white region corresponding to high values of $P_{h_1,h_2}(\mathbf{x})$ favors the first population, while the black region corresponding to low values of $P_{h_1,h_2}(\mathbf{x})$ favors the second population. Once again, the plots give clear decisions for observations 'A' and 'C' but not for 'B'. For observation 'B', the nearly equal spread of white and black regions indicate nearly equal strength of evidence for each of the two populations depending on different choices of the bandwidths. One note-worthy feature of the plots in the two rows in Figure 3.2 is that the plots corresponding to the p-values at the bottom are much sharper than those corresponding to posterior probabilities at the top. Thus the plots in the second row enable an easier visualization of the strength of evidence in favor of one of the two populations for different choices of bandwidths. We next address the reason for such difference in the sharpness for the two sets of plots.

Theorem 3.1 : *For a given observation* $\mathbf{x}$, *assume that* $E\left[K^2\left\{h_i^{-1}(\mathbf{x}-\mathbf{x}_{i1})\right\}\middle|\mathbf{x}\right] < \infty$ *for* $i = 1, 2$. *Define* $\mu_{ih_i}(\mathbf{x}) = E\{\hat{f}_{ih_i}(\mathbf{x})\}$ *for* $i = 1, 2$. *If* $n_1/N \to \lambda$ $(0 < \lambda < 1)$ *as* $N = n_1 + n_2 \to \infty$, *we have*

(a) $\mid \mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x}) - \frac{\pi_1\mu_{1h_1}(\mathbf{x})}{\pi_1\mu_{1h_1}(\mathbf{x})+\pi_2\mu_{2h_2}(\mathbf{x})} \mid = O_P(N^{-1/2})$, *and*

(b) $\mid P_{h_1,h_2}(\mathbf{x}) - I\{\pi_1\mu_{1h_1}(\mathbf{x}) > \pi_2\mu_{2h_2}(\mathbf{x})\} \mid = O_P(N^{-1/2}e^{-CN})$ *for some* $C > 0$, *where* $I\{\cdot\}$ *denotes the indicator function.*

This theorem gives us an idea about the asymptotic behavior of the discrimination measures. For any given $\mathbf{x}$ and a given pair of bandwidths (h_1, h_2), the estimated posterior probability $\mathcal{P}_{h_1,h_2}(\mathbf{x})$ converges to $\pi_1\mu_{1h_1}(\mathbf{x})/[\pi_1\mu_{1h_1}(\mathbf{x}) + \pi_2\mu_{2h_2}(\mathbf{x})]$ at a rate $O(N^{-1/2})$, but depending on $\mu_{1h_1}(\mathbf{x})$, $\mu_{2h2}(\mathbf{x})$ and the prior probabilities, $P_{h_1,h_2}(1 \mid \mathbf{x})$ either converges to 0 or to 1 and that too at an exponential rate. For instance, if $\pi_1\mu_{1h_1}(\mathbf{x}) < \pi_2\mu_{2h_2}(\mathbf{x})$, $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ has a $\sqrt{N}$ rate of convergence to a value less than 0.5, but the corresponding p-value type measure converges to zero at a much faster exponential rate. Therefore, for any given (h_1, h_2), as the training sample size grows, after some stage $P_{h_1,h_2}(\mathbf{x})$ will always give a stronger evidence than $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ for or against population-1.

3.2.3 Misclassification rates

In order to reach a decision regarding classification of an observation $\mathbf{x}$, it is important to know which values of the bandwidth pair (h_1, h_2) is likely to lead to a statistically more reliable classification result. For some fixed choice of (h_1, h_2), consider the average

misclassification probability of a classifier for a two class problem given by

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

where $\mathcal{R}_{h_1,h_2}$ is the set of all $\mathbf{x}$ that are classified into population-1 by the classifier, and $\mathcal{R}^c_{h_1,h_2}$ is the complementary set. Several cross-validation based methods are available in the literature (see e.g., Hills, 1966; Stone [M. Stone], 1977) for estimating the misclassification rate $\Delta(h_1, h_2)$ for a classifier using the training sample. However, as we have pointed out in Section 2.1, these techniques use some kind of empirical proportion of misclassified cases, and a result they lead to estimates that are usually piecewise constant in nature, while the actual function may be a nice smooth function. For varying choices of bandwidths, we have used a smooth and more accurate estimate of the average misclassification probability (as described in Section 2.3) for classifiers, which are based on kernel density estimates. In Figure 3.3, we have plotted that smooth estimate for the average probability of correct classification $\{= 1 - \widehat{\Delta}(h_1, h_2)\}$ using gray scale for the "synthetic data" discussed in the preceding section. Here white color represents high probability of correct classification and black color represents the opposite. This gives a useful visualization of statistical uncertainties in the classification obtained using the kernel density estimates for different levels of smoothing.

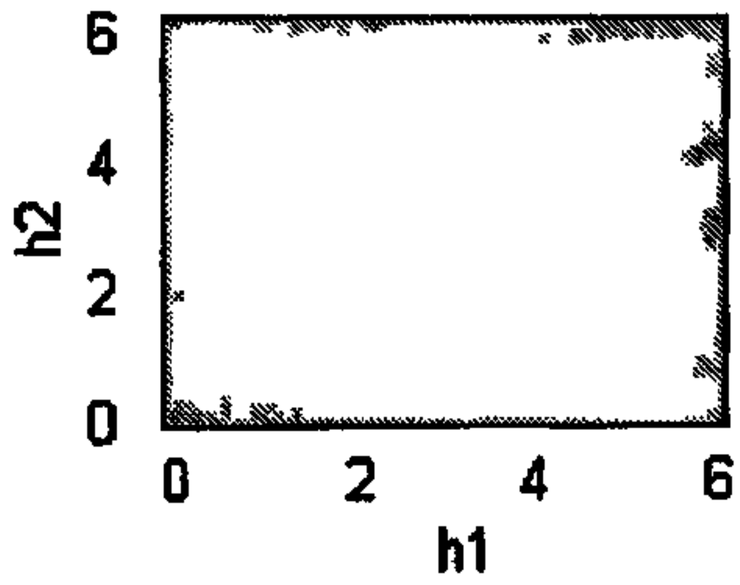


Figure 3.3 : Multi-scale representation for probability of correct classification (synthetic data)

3.3 Aggregation of multi-scale classification results

A natural way to combine the results obtained at different levels of smoothing to arrive at a final decision is to form some appropriate weighted average of the posterior probabilities computed for different choices of (h_1, h_2). We need to use a suitable weight function for this purpose. Clearly, the weight function should take higher values for those pair of bandwidths that lead to lower misclassification rates. There are many reasonable weight functions

satisfying this property. Bagging (see e.g., Breiman, 1996), boosting (see e.g, Freund, 1995; Freund and Schapire, 1997; Schapire *et. al.*, 1998; Friedman, Hastie and Tibshirani, 2000) and arcing classifier (see e.g., Breiman, 1998) are some of the well known methods available in the literature for combining the results of different classifiers to boost up their performance. A comparative study of these ensemble methods can be found in Opitz and Maclin (1999). These methods also assign different weights to different classifiers based on their misclassification probabilities and combine the results using those weights.

3.3.1 Details about the weighted averaging procedure

In this section, we define $\Delta_0 = \min_{h_1,h_2} \widehat{\Delta}(h_1,h_2)$, and consider weight functions $w(h_1,h_2)$ which are decreasing functions of $\widehat{\Delta}(h_1,h_2)$ or equivalently of $\widehat{\Delta}(h_1,h_2) - \Delta_0$. Further, $w(h_1,h_2)$ should vanish whenever the corresponding $\widehat{\Delta}(h_1,h_2)$ exceeds any of the two prior probabilities since the performance of the classifier then turns out to be poorer than that of a trivial classifier, which classifies all observations to the class having the larger prior. Note also that in order to classify an observation $\mathbf{x}$, it would be meaningful to incorporate the corresponding p-value type measure $P_{h_1,h_2}(\mathbf{x})$ in the weights. It makes sense to rely more on those bandwidth pairs, which lead to stronger evidence for one of the two classes and adjust the value of the weight function accordingly. Then, these adjusted weights will not only depend on the estimated overall misclassification probabilities but also on the particular observation to be classified. In all our numerical work, we have used the adjusted weight function

$$w_{\mathbf{x}}(h_1,h_2) = w(h_1,h_2)\,|P_{h1,h2}(\mathbf{x}) - 0.5|\,, \quad \text{where}$$

$$w(h_1,h_2) = \begin{cases} \exp\left\{-\frac{1}{2}\frac{(\widehat{\Delta}(h_1,h_2)-\Delta_0)^2}{\Delta_0(1-\Delta_0)/N}\right\} & \text{if } \frac{\widehat{\Delta}(h_1,h_2)-\Delta_0}{[\Delta_0(1-\Delta_0)/N]^{1/2}} \le \tau \text{ and } \widehat{\Delta}(h_1,h_2) < \min\{\pi_1,\pi_2\} \\ 0 & \text{otherwise.} \end{cases}$$

Here, for $N = n_1 + n_2$, Δ_0 and $\Delta_0(1-\Delta_0)/N$ can be viewed as estimates for the mean and the variance of the empirical misclassification rate of the best classifier based on kernel density estimates, when such a classifier is used to classify N independent observations. The constant τ determines the maximum amount of deviation from the minimal estimated misclassification rate in a standardized scale beyond which the weighting scheme ignores the bandwidth pairs (h_1,h_2) by putting zero weight on them. Clearly, $\tau = 0$ corresponds to the situation of putting all the weights only on the bandwidth pairs (h_1,h_2) for which $\widehat{\Delta}(h_1,h_2) = \Delta_0$. Note also that the choice of the Gaussian-type weight function above implies that for practical purposes there is no need to consider a value of τ larger than 3. This choice of the adjusted weight function is somewhat subjective, and one may use many other suitable functions for the same purpose. However, it is our empirical experience that the final result is not much sensitive to the weighting procedure as long as any reasonable

weight function is used.

We conclude this section by considering once again the synthetic data for the purpose of illustration. In the case of observations 'A' and 'C' (see Section 2.1), the weighted average of the posteriors (with $\tau = 3$) in favor of the first population came out to be 0.81 and 0.24 respectively, which give a clear indication about the classes to which they belong. In the case of observation 'B', however, this weighted average of posteriors in favor of the first population was found to be 0.486, which is very close to 0.5, as one would expect in view of the fact that this observation lies near the class boundary where both the classes have almost equal strength. A more detailed analysis on this data set is given in Section 3.4.

3.3.2 Inadequacy of bandwidths minimizing $MISE$

As we have observed in chapter 2, the bandwidths that are obtained by minimizing $MISE$ of the kernel density estimates, sometimes lead to poor performance in a classification problem. For example, consider the two-class problems discussed in Section 2.1, where both the populations were multivariate normal with the same dispersion matrix $\mathbf{\Sigma} = \mathbf{I}$ but different location parameters $\mu_1 = (2, 0, \ldots, 0)$ and $\mu_2 = (0, 0, \ldots, 0)$. As the distributions themselves were spherical, without any standardization a single common bandwidth was used in all directions. Because of the same dispersion structure of these two populations, it was quite reasonable to use the same bandwidth h for both of them and to consider the average misclassification probability $\Delta(h) = \Delta(h, h)$ as a function of a single bandwidth parameter h. The true $\Delta(h)$-function plotted in Figures 2.1 and 2.3 showed the difference between the optimal bandwidth for minimizing the $MISE$ in kernel density estimation (marked by 'o') and that for minimizing the average misclassification probability (marked by '*'). In almost all the cases, h_* led to a significantly lower misclassification error rate than that obtained by using h_o.

Here, we carry out a simulation study with equal number of observations from the two classes $N_6(\mu_1, \mathrm{I}_6)$ and $N_6(\mu_2, \mathrm{I}_6)$. We generate a test set of size 1000 (500 from each class) and classify them using 100 training set observations (50 from each class). The classifier that uses the bandwidth minimizing the estimated $MISE$ of kernel density estimates misclassifies 223 observations, whereas the optimal Bayes classifier wrongly classifies only 162 observations. When weighted averages of the posteriors are used, the number of misclassified cases turns out to be 185 and 191 respectively for $\tau = 0$ and $\tau = 3$.

However, it is important to notice that the use of h_* (which is equivalent to the method of weighted averaging of posteriors with $\tau = 0$) does not necessarily lead to better estimates for the posterior probabilities. In Figure 3.4, the estimated posterior probabilities for the simulated data discussed above are plotted against the true posteriors of different

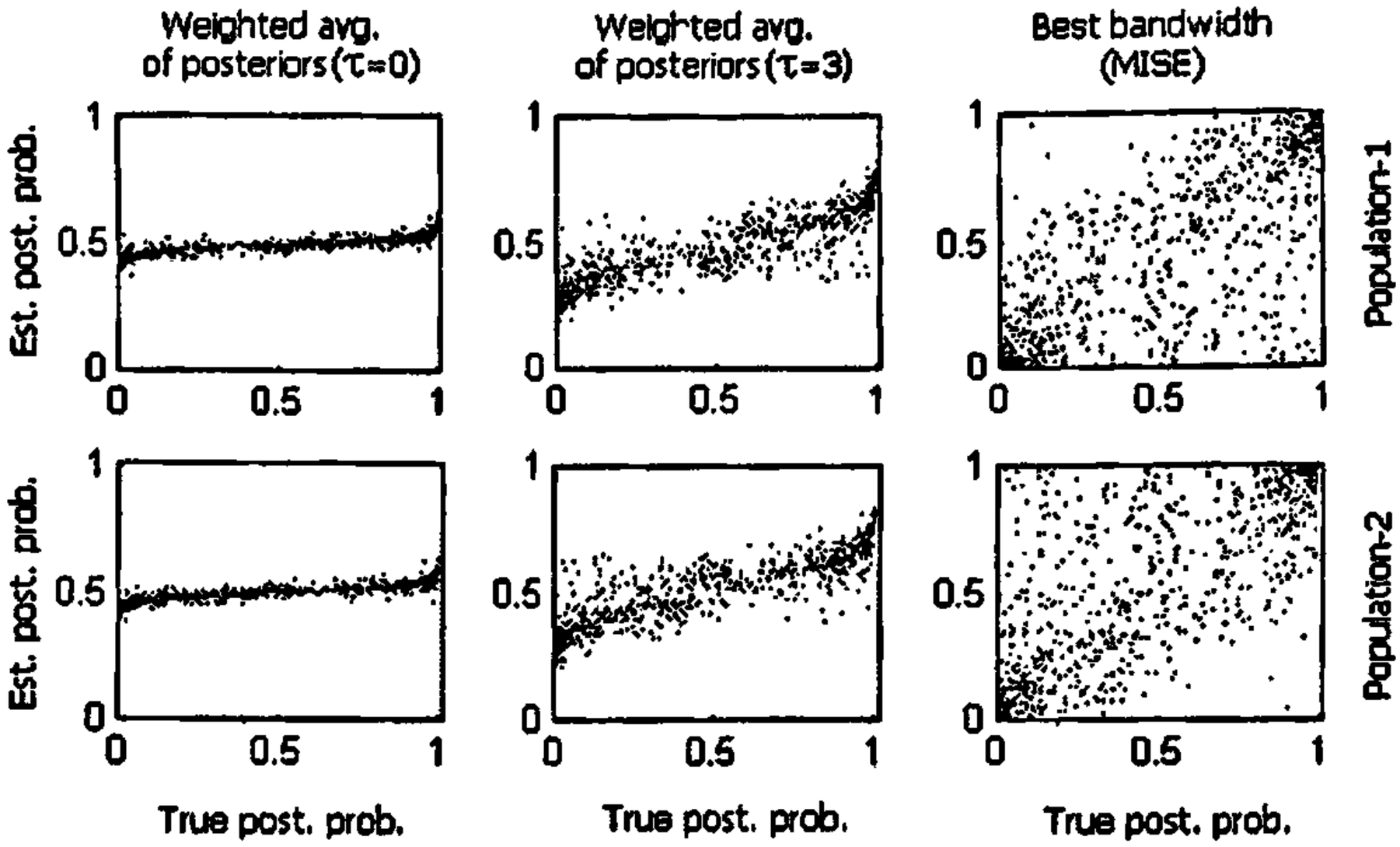


Figure 3.4 : Estimated posterior probabilities for simulated data set

observations. When h_o is used as the common bandwidth (right column of Figure 3.4), the posteriors get more scattered but this choice of bandwidth leads to very little bias for the posterior probability estimates. On the other hand, when h_* is used (left column of Figure 3.4), the scatter shrinks to the horizontal line at the center indicating a reduction in variance of the estimates but the bias of the posterior estimates increases considerably. In Figures 2.1 and 2.3, we have observed h_* to be much larger than h_o. This larger bandwidth reduces the variance of the kernel density estimate at the cost of increased bias, and this is precisely the fact that is reflected in Figure 3.4. While h_o leads to a mean square error of 0.046 for posterior estimates, h_* increases it to 0.112. The method based on weighted averaging of posterior with $\tau = 3$ amounts to a compromise between the preceding two. It improves the mean square error (0.061) of posterior estimates significantly without sacrificing much accuracy in terms of misclassification rates. A detailed discussion on the effect of such bias and variance on misclassification error rates is available in Friedman (1997).

3.3.3 Classification among more than two populations

When there are more than two classes, due to high computational complexities, it is difficult to find out the optimum bandwidths that minimize some estimates of overall average misclassification probability $\Delta(h_1, h_2, \ldots, h_J)$. In these situations, we can decompose the multi-class problems into a number of binary classification problems and proceed in the same way as before. The results of all these two-class problems are combined together to come up with the final decision rule. The method of majority voting (see e.g., Friedman, 1996) is the simplest procedure for combining the results of pairwise classifications. In a J-class

problem, after $\binom{J}{2}$ pairwise comparisons, this method classifies an observation to the class which has the maximum number of votes. However, this voting method sometimes may lead to a region of indecision, where more than one class can have the maximum number of votes. One can avoid this problem using alternative techniques like the method of pairwise coupling (see e.g., Hastie and Tibshirani, 1998), which combines the estimated posteriors for different pairwise classifications to determine the final posteriors for the competing classes.

We conclude this section by pointing out that we have observed the inadequacy of bandwidths that minimize $MISE$ of population density estimates in discriminant analysis of some real data sets in addition to the simulated example reported earlier. For instance, consider the "diabetes data" involving five measurement variables (fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight) and three classes of individuals ("overt diabetic", "chemical diabetic" and "normal") reported in Reaven and Miller (1979). There are 145 individuals with 33, 36 and 76 in the three classes according to some clinical classification. For this data set, if we use bandwidths that minimize estimated $MISE$ of population density estimates, we get leave-one-out cross-validated misclassification rate of 12.41%. This error rate is higher than that obtained for even simple linear and quadratic discriminant analysis, which showed leave-one-out misclassification rates of 11.03% and 9.66% respectively. For k-nearest neighbor method this error rate was found to be 8.97% when k was chosen by cross-validation technique. However, for our multi-scale analysis followed by the weighted averaging of posteriors led to a leave-one-out cross-validated error rate of 6.21% for both $\tau = 0$ and $\tau = 3$. Note that this is a three class problem, and we have used the method of "majority voting" to combine the results of pairwise comparisons to arrive at the final classification. Fortunately, in this data set, majority voting did not lead to any tied case either for $\tau = 0$ or for $\tau = 3$.

3.4 Case studies using benchmark data sets

In this section, we report our findings based on some benchmark data sets that illustrate the utility of the proposed method. Results of the kernel discriminant analysis based on bandwidths that minimize $MISE$ and that based on the weighted averaging of posteriors are presented to compare their performance. For classification problems with more than two populations, we adopt the pairwise classification method and combined the results by using majority voting (Friedman, 1996) as well as by pairwise coupling (Hastie and Tibshirani, 1998). Misclassification error rates for usual linear and quadratic discriminant analysis (LDA and QDA) are also given to facilitate the comparison. As we have discussed earlier, in a few cases, the voting method may end up with a tied situation. Here, all those tied cases are considered as "misclassification". Therefore, the reported results on voting are

actually the proportion of misclassifications in the worst possible cases. The data sets we consider here have been analyzed before in the literature, where nonparametric methods like classification trees, neural nets and flexible discriminant analysis (FDA) based on multivariate adaptive regression splines ($MARS$, see Friedman, 1991) was used to classify the observations. We have quoted those results directly from the available literature, and they are also presented in Table 3.1. Because of the unavailability of error rate for FDA on image segmentation data we have a blank space in the table. Throughout these experiments, sample proportions for different classes are used as their prior probabilities. Apart from the vowel recognition data-1, all the data sets that are considered in this section are available at http://www.lib.stat.cmu.edu or at http://www.uci.ics.edu.

Synthetic data : Description of this data set has already been given in Section 3.2. Ripley (1994) used this data to compare the performance of different classification algorithms. The class distributions were chosen to have a Bayes risk of 8.0%. In this data set, LDA and QDA could achieve test set error rates of 10.8% and 10.2%, respectively. Classification tree ($CART$) (see e.g., Breiman *et. al.*, 1984) also misclassified more than 10% observations (see Table 3.1). Performance of other nonparametric methods was fairly similar. Weighted averaging of the posterior achieved the best error rate when $\tau = 0$ is used. Since it is a two-class problem, coupling and voting led to the same result.

Vowel recognition data -1 : This data was created by Peterson and Barney (1952) by a spectrographic analysis of vowels in words formed by 'h' followed by a vowel and then followed by 'd'. There were 67 persons who spoke different words and the two lowest resonant frequencies of a speaker's vocal track were noted for 10 different vowels. The observations were then randomly divided into a training set consisting of 338 observations and a test set consisting of 333 observations. Here, the classes have significant overlaps between them, which makes the data set a challenging one for any classification method. A scatter plot of this data set is given in Figure 3.5, where the numbers represent the labels of different classes ('0' represents the 10-th class).

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

When pairwise coupling was used for final classification, we obtained an error rate

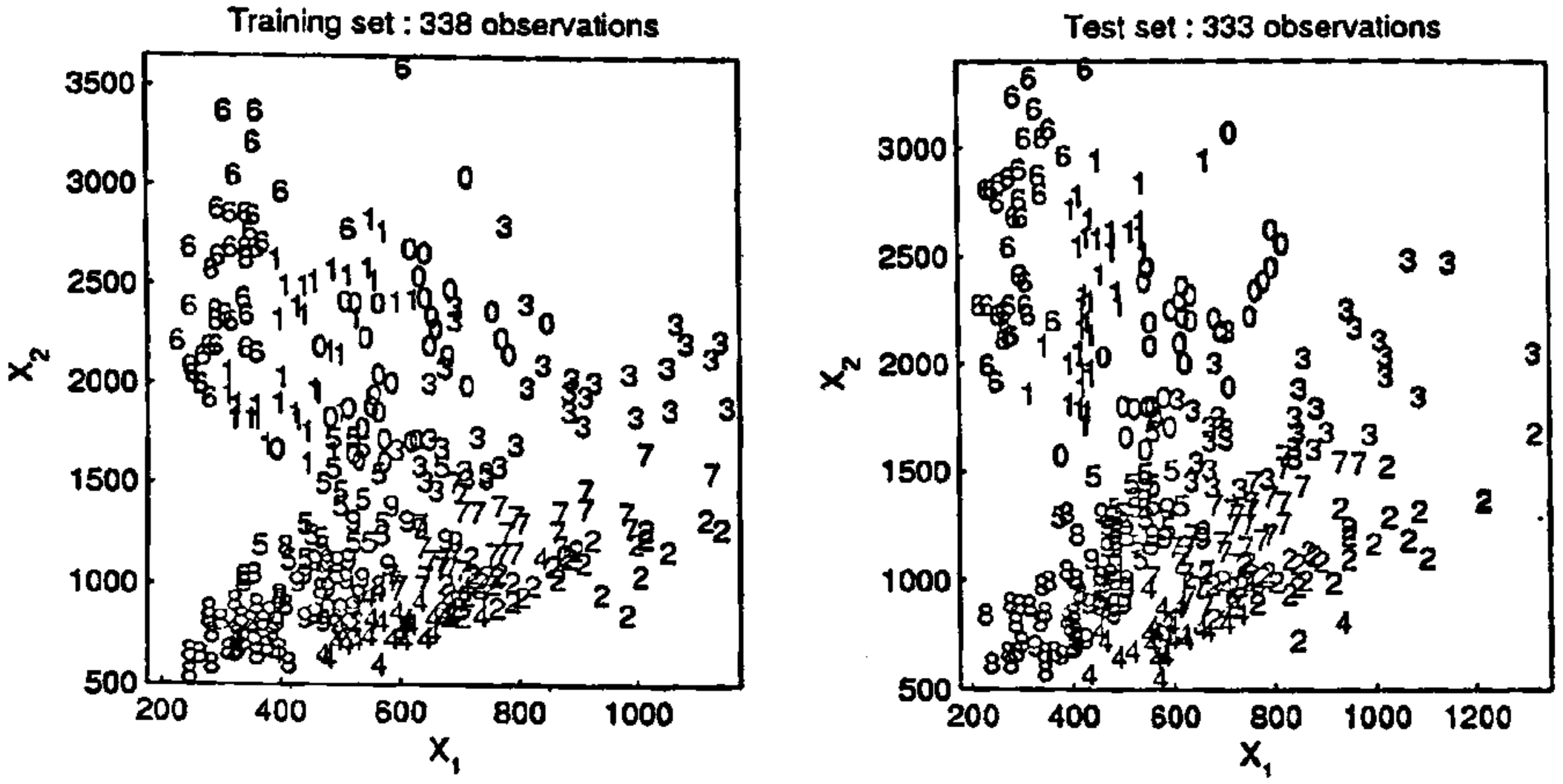


Figure 3.5 : Scatter plots for vowel recognition data-1

of 36.4% for weighted averaging of posteriors with $\tau = 0$. We suspect that since the optimal bandwidth minimizing the misclassification rate does not always lead to good estimates for posterior probabilities as we have seen before, the performance of pairwise coupling method turns out to be so bad due to the presence of a large number of overlapping populations. The posterior estimates are expected to become better when $\tau = 3$ is used instead of $\tau = 0$. Perhaps this could be one of the reason for improved performance of the classifier leading to an error rate of 22.8% when we used weighted averaging of posteriors with $\tau = 3$ followed by an application of pairwise coupling method.

Vowel recognition data -2 : This data set is also related to a vowel recognition problem where ten measurements on speech signal are taken on each observation corresponds to one of the 11 vowels (see Robinson (1989) for detail description of this data set). There are 528 observations in the training set and 462 observations in the test set equally distributed among these 11 competing classes. This is a difficult data set and a part of this difficulty arises due to the presence of a fairly large number of competing populations. In this data set, many well-known classifiers misclassified more than 50% of the test set observations. The best error rate (42%) was observed for $FDA - MARS$ (degree 2). Error rates for the multi-scale methods were quite competitive as compared to the other nonparametric classifiers.

Sonar data : This data set was used by Gorman and Sejnowski (1988). It contains 111 patterns obtained by bouncing sonar signals off a metal cylinder and 97 patterns obtained from rocks at various angles and under various conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. Signals were obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for

the rock. Each observation is a set of 60 numbers in the range 0.0 to 1.0, each of which represents the energy within a particular frequency band, integrated over a certain period of time. To reduce co-ordinate-wise dependence, the data were averaged in a band of three making the number of measurement variables 20. The data set was split into a training and a test set of size 104 each using a cluster analysis method to ensure even matching.

Results for different classification methods on this data set are available in Ripley (1994) and Coolie and MacEachern (1998). In this data set, QDA performed quite well as compared to other classification methods like LDA, $FDA - MARS$, $CART$ and neural nets (see Table 3.1). However, the kernel methods performed even better with the best result obtained for weighted averaging of posterior with $\tau = 3$.

Image segmentation data : This data set contains 19 different measurements on each image of one of the seven different objects. There are 210 observations in the training and 2100 observations in the test set which are equally distributed in those 7 classes. The data set and the description of the variables are available at http://www.uci.ics.edu. The value of the 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. Among different classifiers, weighted averaging methods and LDA had better misclassification rates.

Classification methods	Synthetic	Vowel-1	Vowel-2	Sonar	Image
LDA	10.8	25.2	55.6	20.2	11.4
QDA	10.2	19.8	52.8	15.4	14.6
$FDA - MARS$	9.3	20.7	45.0	22.1	*
(degree 2)	9.6	19.8	42.0	19.2	*
Classification tree	10.1	23.7	56.4	20.2	12.6
Neural networks	9.4	18.6	50.9	19.2	12.1
Kernel ($MISE$)	9.3	18.9	62.1	14.4	15.7
Kernel (Wt. avg.)					
Voting ($\tau = 0$)	9.0	19.8	50.6	14.4	10.5
Coupling($\tau = 0$)	9.0	36.6	47.2	14.4	11.7
Voting ($\tau = 3$)	9.2	17.4	51.9	12.5	11.0
Coupling ($\tau = 3$)	9.2	22.8	48.9	12.5	10.6

* Results are not available

Table 3.1 : Misclassification rates (in %) for different classification methods

3.4.1 Glass data : a challenging problem for kernel classifiers

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

Ripley (1996) analyzed this data set extensively and reported cross-validated error rates for different classifiers. The best result was reported for the k nearest neighbor method with $k = 1$, when the measurement variables were suitably re-scaled. This re-scaled nearest neighbor algorithm had an error rate of 23.6%. Misclassification error rate for usual nearest neighbor method was found to be 26.6% for $k = 1$. Neural networks with 4 to 8 hidden nodes were reported to have error rates between 24.8% and 29.9%. LDA in this data set led to a cross-validated error rate of 37.9%, which was even worse for quadratic discriminant analysis (error rate = 40.2%). Classification tree methods had error rates ranging from 31% to 42% for different types of pruning. $FDA-MARS$ (with degree 1) could achieve the error rates of 32.2%, which was reduced to 29% when interactions were taken into consideration. Logistic discriminant analysis (see e.g., Ripley, 1996; Hastie *et. al.*, 2001) and projection pursuit (see e.g., Huber, 1985) had higher error rates (36% and 35.5% respectively) than the other nonparametric classifiers.

As four out of the nine measurement variables (oxides of Mg, K, Ba and Fe) have a significant number of zeros among their observed values, we decided to carry out our analysis with the remaining five variables. However, even after using this subset of measurement variables, we could achieve a competitive performance for classifiers based on kernel density estimates. When the bandwidths, which minimize $MISE$ of the density estimates, are used for classification, they led to a fairly good performance. The leave-one-out estimate for the misclassification rate was found to be 31.3%. Using the method of weighted averaging of posterior, we obtained even better performance. The error rates for $\tau = 0$ and $\tau = 3$ were found to be 29.9% and 28% respectively when majority voting is used. When pairwise coupling method was applied to this data set after weighted averaging of posteriors, the aforesaid error rates increased to 32.7% and 34.6% respectively.

3.5 Behavior of $\Delta(h_1, h_2)$ in equal and unequal prior cases

As we have seen earlier, for some fixed h_j, $\hat{f}_{jh_j}(\mathbf{x})$ is an average of i.i.d. random variables. When the sample size n_j gets larger, its variance shrinks to 0, and it converges in probability to $\mu_{jh_j}(\mathbf{x}) = E\{\hat{f}_{jh_j}(\mathbf{x})\} = K_{h_j} * f_j(\mathbf{x})$, which is a convolution of the density function f_j with a kernel K with bandwidth h_j. We know that in a two-class problem, the ordering of $\pi_1\mu_{1h_1}(\mathbf{x})$ and $\pi_2\mu_{2h_2}(\mathbf{x})$ determines the asymptotic decision rule. Therefore, in equal prior cases, asymptotic classification depends on the ordering of $\mu_{1h_1}(\mathbf{x})$ and $\mu_{2h_2}(\mathbf{x})$. When the K and the f both are spherically symmetric, and they are decreasing in distance from their centers of symmetry, the convolution $K_h * f$ also retains that property for every positive value of h (see Section 2.2). Further, if the population distributions satisfy the location shift model [i.e. $f_j(\mathbf{x}) = g(\mathbf{x} - \mu_j)$ for some common density g and location parameters μ_j], and same bandwidth h is used for both populations, $\mu_{jh}(\mathbf{x})$ $(j = 1, 2)$ preserves the ordering of the actual densities for all values of h. Therefore, in equal prior cases, for all positive values of $h = h_1 = h_2$, the corresponding misclassification probability asymptotically turns out to be the optimal Bayes risk. For large h, as the variance tends to zero rather quickly, this convergence is much faster. So, in this case, one expects to have high values of $1 - \widehat{\Delta}(h_1, h_2)$ along the diagonal $h_1 = h_2$ in the plot of the probability of correct classification. The following theorem gives some idea about the behavior of the classifier based on kernel density estimates for large sample sizes and large bandwidths in general situation when the population densities may not necessarily satisfy any symmetry condition.

Theorem 3.2 : *Suppose that f_1 and f_2 are such that $\int \|\mathbf{x}\|^6 f_i(\mathbf{x})d\mathbf{x} < \infty$ and the kernel K is a d-dimensional density function with a mode at 0 and bounded third derivatives. Define a constant $C_\pi = \pi_2/\pi_1$ and assume that h_1, h_2 vary in such a way that $h_2/h_1 = C_h$, a constant. Now as $h_1 \to \infty$, $\Delta(h_1, h_2)$ has the following asymptotic behavior.*

(a) When $C_\pi > C_h^d$, as $n_1, n_2 \to \infty$, $\Delta(h_1, h_2) \to \pi_1$.

(b) When $C_\pi < C_h^d$, as $n_1, n_2 \to \infty$, $\Delta(h_1, h_2) \to \pi_2$.

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

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

Note that when $C_\pi = C_h = 1$ (i.e., $\pi_1 = \pi_2$), the above quadratic classifier actually turns out to be a linear classifier given by

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

Under the assumptions on location shift and spherical symmetry (as discussed above), and when the kernel function K is also spherical (note that $\nabla^2 K(0)$ is negative definite), this

linear classifier can be expressed in a further simplified form as

$$\mathrm{d}_l(\mathbf{x}) = \arg\max_i \left\{ \mathbf{x}^{'} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^{'} \boldsymbol{\mu}_i \right\},$$

where $\boldsymbol{\mu}_i$ is the location parameter for the i-th population ($i = 1, 2$). It is to be noted that the linear classifier described above is the optimal Bayes classifier under this set up. Therefore, in this particular case, the misclassification probability Δ asymptotically converges to the optimal Bayes risk.

To illustrate the above findings, we consider a simple two-class problem in two-dimension, where both the populations are normally distributed. A sample of size 100 is generated from each of the two populations [$N_2(0, 0, 1, 1, 0)$ and $N_2(2, 0, 1, 1, 0)$] to form our training sample. Estimated misclassification probabilities are computed over a wide range of bandwidths, and the corresponding probabilities of correct classification are plotted in Figure 3.6. When the priors are equal, for large values of $h_1 (= h_2)$, the points on the diagonal represents approximately the performance of the usual linear discrimination rule, which is the optimal Bayes classification rule here. This is reflected in Figure 3.6(a) where we observe a white strip [high values of $1 - \hat{\Delta}(h_1, h_2)$] along the diagonal indicating a good performance for the kernel density estimate based classifier. But for the other large values of bandwidths (where $h_1 \neq h_2$), $\hat{\Delta}(h_1, h_2)$ becomes close to 0.5.

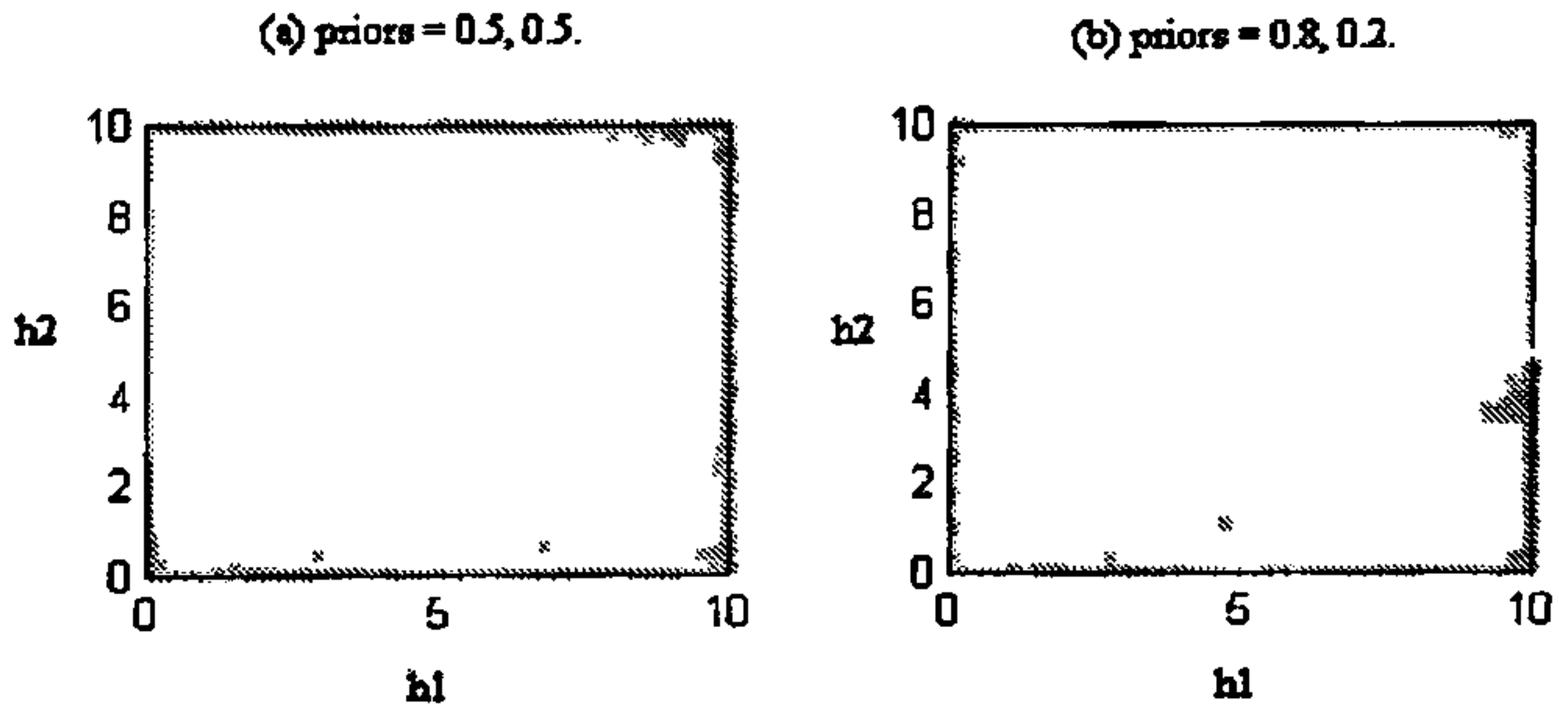


Figure 3.6 : Estimated probabilities of correct classification (two-dimensional simulated data)

However, the plot gets completely changed [see Figure 3.6(b)] if we choose the priors to be unequal. An example of that type is considered here, where the observations for the two normal populations are kept unchanged but the priors are set as 0.8 and 0.2 ($C_\pi = 0.25$), respectively, for the two classes. From Theorem 3.2, we know that, along the line $h_1 = 2h_2$ ($C_h = 0.5$), for large values of the sample size and the bandwidth parameters, $\hat{\Delta}$ converges to the misclassification probability of a quadratic classifier, which is nearly 16% for this problem. For this reason, we observe a narrow white strip along the line $h_1 = 2h_2$

but unlike the equal prior case, the patterns on the two sides of the line are completely different. For the pairs of bandwidths above the line (i.e., when $2h_2 > h_1$), we know $\widehat{\Delta}$ tends to $\pi_2 = 0.2$ but for the pairs below the line (i.e. when $2h_2 < h_1$), it converges to $\pi_1 = 0.8$. This is why a darker region is observed at the right hand bottom corner. These plots show that not only the population density functions, but also the priors have a major role in determining the behavior of smoothing parameters in kernel discriminant analysis.

3.6 Remarks and discussions

Our multi-scale method is a more informative approach than the usual kernel discriminant analysis, where along with the classification results we get an idea about the strength of the discrimination measure and related statistical uncertainties present there. In a method based on single optimum bandwidth, that bandwidth does not depend on the specific observation to be classified. The use of data dependent adjusted weight function in our multi-scale method provides that flexibility, which is very desirable.

In a multi-class problem, often it is computationally difficult to find out the optimal bandwidths by minimizing the estimated misclassification probability $\widehat{\Delta}(h_1, h_2, \ldots, h_J)$. The pairwise classification method used in this chapter not only reduces the computational burden significantly, but also provides the flexibility of using different bandwidths for a class when we compare it to different competing classes. Since a good choice of the bandwidth for a population density estimate in a classification problem may depend on its competing population density estimates as well as the corresponding prior probabilities, this pairwise comparison approach seems to be quite meaningful.

One of the main objective behind the multi-scale method is to provide a useful device for easier visualization of the strength of classification at different levels of smoothing, which may not be obtained in a single bandwidth based classification method. For example, consider the six-dimensional simulated data (as discussed in Section 3.3.2) and also consider an observation $\mathbf{x} = (x_1, 0, 0, 0, 0, 0)$. Clearly, $x_1 = 0$ and $x_1 = 2$ give the center for population-2 and population-1, respectively, while $x_1 = 1$ represents a point on the class boundary. Therefore, one normally expects to have three different behavior of the classification methodology at these three points. When h_* is used as the bandwidth parameter, the posterior probabilities for $\mathbf{x}$ are observed to be 0.471, 0.499 and 0.528 respectively for $x_1 = 0$, 1 and 2, from which the difference in the strength of classification is not very clear. In Figure 3.7, the posterior probabilities and p-values for these observations are plotted for various choices of bandwidth parameters along with the corresponding probabilities of correct classification. Adjusted weights (re-scaled to have maximum value 1 and minimum value 0) are also presented in the last column of the figure, where white color points out

the regions of high reliability. In the plots of both posterior probabilities and p-values, the white region extends as we move on from $x_1 = 0$ to $x_1 = 2$. However, the strength of the decision is not quite clear from these figures since in all the cases we have almost equal split in favor of the classes indicated by white and black regions.

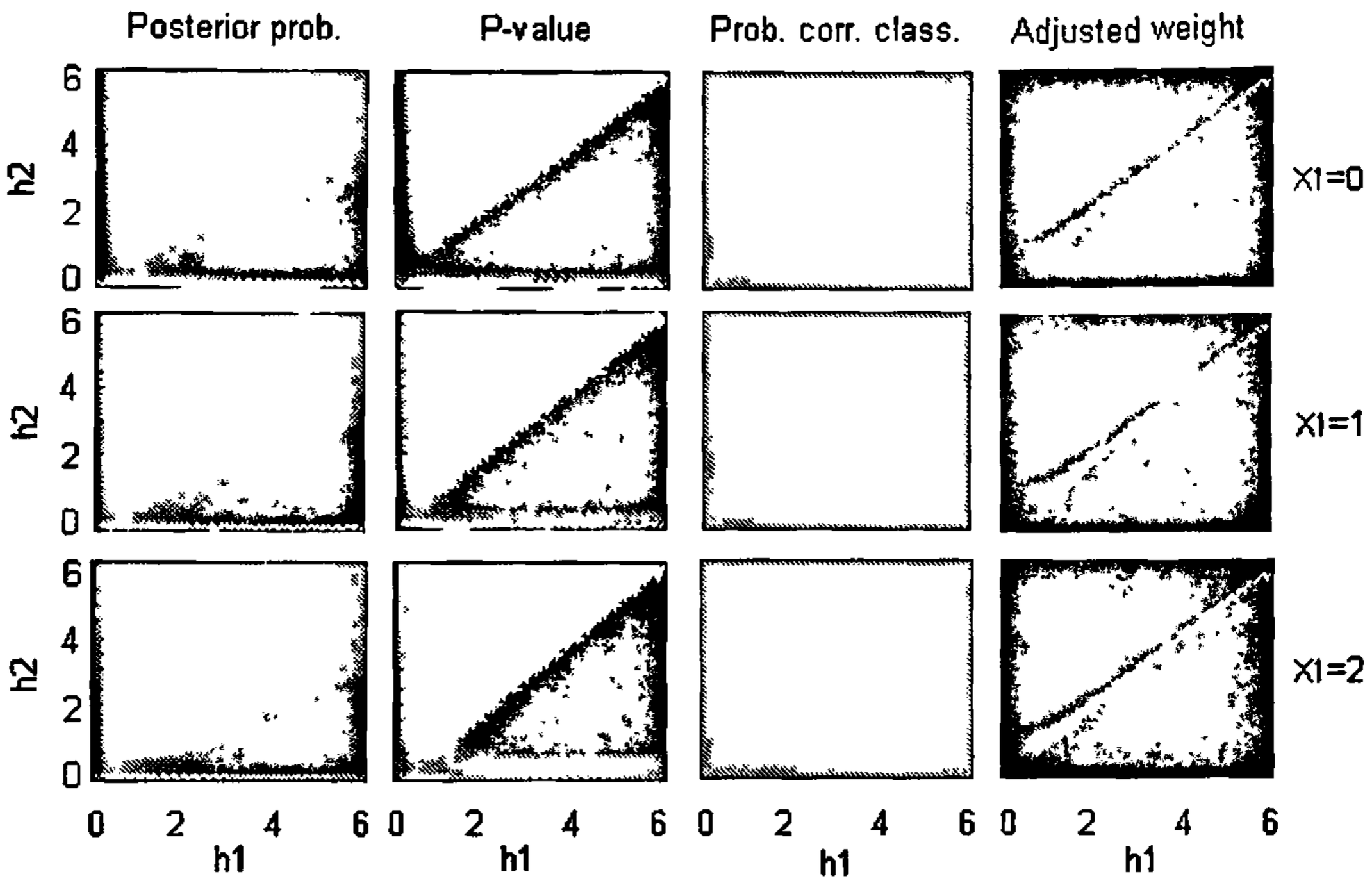


Figure 3.7 : Results of multi-scale analysis on simulated data

But the difference in the classification results and that in the strength of discrimination become evident if we look at the adjusted weight function with the sign same as that of $P_{h_1,h_2}(\mathbf{x}) - 0.5$. In that case, if we re-scale the signed weight function to have a maximum value of 1 and minimum value of 0, the re-scaled version can be expressed as $W^*_{\mathbf{X}}(h_1, h_2) = 0.5 + Sign\{P_{h_1,h_2}(\mathbf{x}) - 0.5\} w^*_{\mathbf{X}}(h_1, h_2)$, where $w^*_{\mathbf{X}}$ is the re-scaled version of the adjusted weight function plotted in the last column of Figure 3.7. $W^*_{\mathbf{X}}(h_1, h_2)$ can also be viewed as a super-imposition of the plots of p-values over that of the weight functions $w(h_1, h_2)$.

Figure 3.8 shows the importance of this signed weight function in visual representation of the strength of discrimination of a kernel density estimate based classifier. It is quite clear from the definition that when the pair (h_1, h_2) has low weight, $W^*_{\mathbf{X}}(h_1, h_2)$ is expected to be very close to 0.5, which is indicated by the gray regions in the plots. However, in more reliable regions (pairs having high weights), we get stronger evidence as $P_{h_1,h_2}(\mathbf{x})$ moves away (in either direction) from 0.5. When $x_1 = 0$ (or $x_1 = 2$), we observe a black (or white) shade in this region, which gives a clear idea about the direction and the strength of the decision. Evidence for classification is very strong in these cases. For $x_1 = 1$, we observe

some white as well as some black shades of almost equal intensity. Clearly, the evidence is poor in this case, and the plot gives a clear indication of a border line case.

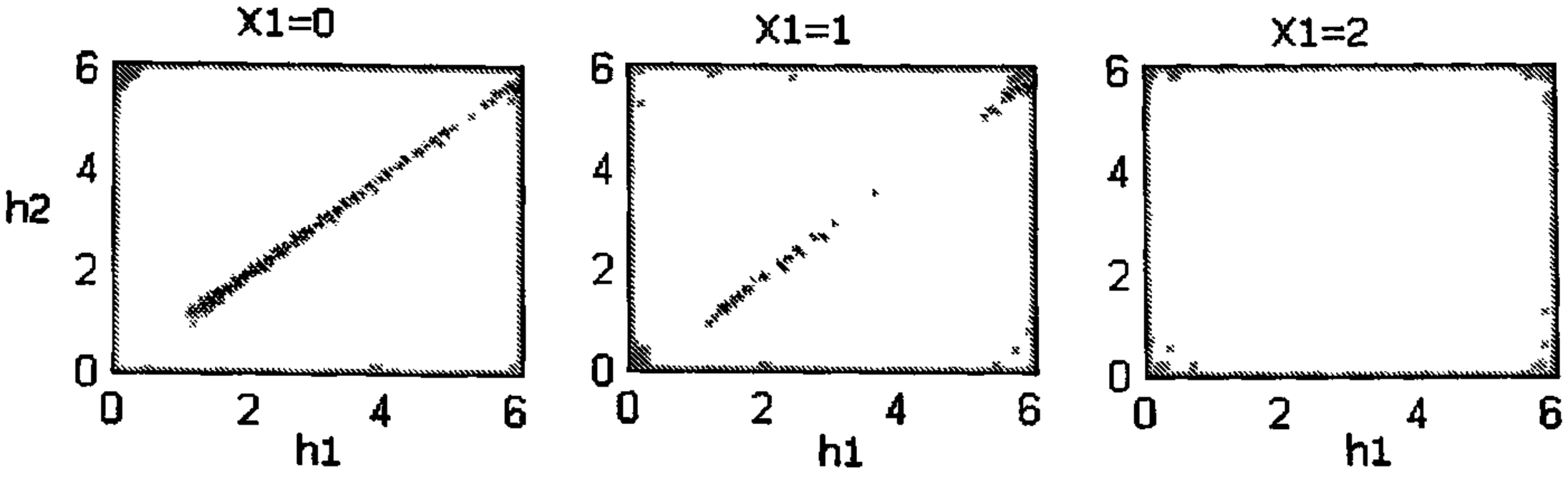


Figure 3.8 : Plots for signed adjusted weight function after re-scaling

In the plots of posterior probabilities and p-values, one may notice a white or a black streak near each of the two axes. This is because for the given sample sizes, use of very small bandwidth makes one density estimate very close to zero and therefore the competing class density estimate turns out to be the winner. However, these streaks appear in a region of the plot where misclassification rates are high (see third column of Figure 3.7). Consequently, the weight function becomes almost zero in those regions, and the aggregation procedure does not get affected.

3.7 Proofs and mathematical details

Proof of Theorem 3.1 : (*a*) To make the expressions notationally simpler, let us define $T_i = \pi_i \hat{f}_{ih_i}(\mathbf{x})$ for $i = 1, 2$. Now, as T_i is an average of i.i.d. random variables, from Central Limit Theorem, it follows that under the assumed moment condition, for large sample sizes, T_i tends to be normally distributed with mean $\tau_i = \pi_i \mu_{ih_i}(\mathbf{x})$ and variance $v_i = \pi_i^2 s^2_{ih_i}(\mathbf{x})$, which is of the order $O(n_i^{-1})$.

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

$$\frac{\{\Psi(T_1, T_2) - \Psi(\tau_1, \tau_2)\}}{\mathcal{V}} \xrightarrow{L} Normal\ (0, 1), \quad \text{where } \mathcal{V} = \left\{ \sum_{i=1}^{2} v_i \, (\partial \Psi / \partial T_i)^2_{T_1=\tau_1, T_2=\tau_2} \right\}^{1/2}$$

Since $n_1/N \to \lambda$ and $n_2/N \to 1-\lambda$ $(0 < \lambda < 1)$ as $N \to \infty$, we have $|\Psi(T_1, T_2) - \Psi(\tau_1, \tau_2)| = O_P(N^{-1/2})$.

(*b*) Without loss of generality, let us assume that $\tau_1 > \tau_2$ i.e. $I\{\pi_1 \mu_{1h_1}(\mathbf{x}) > \pi_2 \mu_{2h_2}(\mathbf{x})\} = 1$.

Now, for some fixed h_1, h_2 and $\mathbf{x}$, from part (a) of this theorem, it follows that,

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

Now define $Z_{h_1,h_2}(\mathbf{x}) = \frac{1}{\sqrt{v_1+v_2}}(T_1-T_2)$. From the order of v_1 and v_2 it is easy to see that $Z_{h_1,h_2}(\mathbf{x}) = O_p(N^{1/2})$, and $\frac{1}{\sqrt{N}}Z_{h_1,h_2}(\mathbf{x})$ converges to a constant (C, say) in probability. Therefore, for $x>0$, using the fact that $\frac{1}{x}\phi(x) < 1-\Phi(x) < \left(\frac{1}{x}-\frac{1}{x^3}\right)\phi(x)$, (where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the p.d.f. and the c.d.f. of a standard normal distribution, respectively), we get $1-P_{h_1,h_2}(\mathbf{x}) = 1-\Phi\left(Z_{h_1,h_2}(\mathbf{x})\right) = O_p(N^{-1/2}e^{-CN})$.

Proof of Theorem 3.2 : First note that

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

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

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

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

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

where $Y_{i,k,l} = (x_i-X_i)(x_k-X_k)(x_l-X_l)\frac{\partial^3K(\mathbf{t})}{\partial t_i\partial t_k\partial t_l}|_{\mathbf{t}=\xi}$ for some intermediate vector ξ between $\mathbf{0}$ and $(\mathbf{x}-\mathbf{X})/h_j$. Therefore, using the fact that K has bounded third derivatives and $\int\|\mathbf{x}\|^6 f(\mathbf{x})d\mathbf{x} < \infty$, we obtain

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

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

$$\begin{aligned} &\pi_1E_{f_1}\{\hat{f}_{1h_1}(\mathbf{x})\} > \pi_2E_{f_2}\{\hat{f}_{2h_2}(\mathbf{x})\} \\ &\Leftrightarrow \pi_1h_1^{-d}\left[K(0)+(1/2h_1^2)E_{f_1}\{(\mathbf{x}-\mathbf{X})'\nabla^2K(\mathbf{0})(\mathbf{x}-\mathbf{X})\}+O(h_1^{-3})\right] \\ &\qquad > \pi_2h_2^{-d}\left[K(0)+(1/2h_2^2)E_{f_2}\{(\mathbf{x}-\mathbf{X})'\nabla^2K(\mathbf{0})(\mathbf{x}-\mathbf{X})\}+O(h_2^{-3})\right] \\ &\Leftrightarrow C_\pi C_h^{-d}\left[K(0)+(1/2h_1^2)E_{f_1}\{(\mathbf{x}-\mathbf{X})'\nabla^2K(\mathbf{0})(\mathbf{x}-\mathbf{X})\}+O(h_1^{-3})\right] \\ &\qquad > \left[K(0)+(1/2h_2^2)E_{f_2}\{(\mathbf{x}-\mathbf{X})'\nabla^2K(\mathbf{0})(\mathbf{x}-\mathbf{X})\}+O(h_2^{-3})\right]. \end{aligned}$$

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

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

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

Chapter 4

Visualization and aggregation of nearest neighbor classifiers

4.1 Main problem and motivation

Like kernel discriminant analysis, nearest neighbor classification is also a very popular nonparametric discriminant analysis procedure, and we are going to investigate that in this chapter. In order to classify an observation by k-nearest neighbor method (k-NN) (see e.g., Fix and Hodges, 1951; Cover and Hart, 1968; Devijver and Kittler, 1982; Dasarthy, 1991), we assume that the posterior probability of a specific class to be constant over a small neighborhood around that observation. Generally, a closed ball of radius r_k is taken as this neighborhood, where r_k is the distance between the observation and its k^{th} nearest neighbor. We classify an observation to the class which has the maximum number of representatives in this neighborhood. The parameter k, which determines the volume of this neighborhood, can be viewed as a "smoothing parameter" that is related to the smoothness of the posterior conditional probability, and in future we will refer to it as the *neighborhood parameter*. As k gets larger, the posterior probability estimates tend to be smoother in some sense. A discussion on the bias and the variance of the posterior probability estimates for different k's is available in Fukunaga and Hostetler (1973) and Friedman (1997). The performance of the nearest neighbor classification rule depends heavily on the value of this neighborhood parameter. Existing theoretical results (see e.g., Loftsgaarden and Quesenberry 1965; Cover and Hart, 1968; Fukunaga and Hostetler, 1973; Stone, 1977; Devroye, 1981) suggest that k should depend on the training sample size N, and it should vary with N in such a way that $k \to \infty$ and $k/N \to 0$ as $N \to \infty$. However, in practice, the optimal value of k depends on the available training sample observations, and one generally uses re-sampling techniques like cross-validation (see e.g., Hills 1966; Stone [M. Stone], 1977) to find it

out. One should also notice that the optimal value of k is case specific, and it depends on the observations to be classified in addition to depending on the competing population distributions. Further, for a specific observation, one may also like to assess the strength of evidence for different classes for varying choices of k. Therefore, in classification, instead of relying on a single value of k, it will be more useful to look at the results for different k's to come up with the final decision. This chapter presents one such approach, where the results for different neighborhood parameters are studied simultaneously in order to build up a more informative classification procedure. These results are presented in a two dimensional plot, which is specific to an observation to be classified. In this plot, one can visually compare the strength of different classes at some particular region of the sample space. In Chapter 3, we have discussed about such a visual approach for kernel discriminant analysis, where a range of values have been used for the bandwidth parameters of the kernel density estimates of different competing classes. Chaudhuri and Marron (1999, 2000) used similar ideas to extract the features in a function estimation problem. Uncertainty of different classification rules (for varying choices of neighborhood parameter) are judged on the basis of corresponding estimated misclassification probabilities, which are also presented in another two dimensional plot. Like multi-scale kernel discriminant analysis discussed in Chapter 3, all these informations are combined together to arrive at the final result.

4.2 Description of the methodology

Given an observation $\mathbf{x}$, let $\mathbf{x}^{(k,N)}$ be the its k^{th} nearest neighbor and $r_k = \rho(\mathbf{x}, \mathbf{x}^{(k,N)})$ be the distance between them. A k-nearest neighbor rule classifies the observation $\mathbf{x}$ to the j^{th} population if $\sum_{k=1}^{N} I\{\rho(\mathbf{x}, \mathbf{x}_k) \leq r_k,\ c_k = j\} \geq \sum_{k=1}^{N} I\{\rho(\mathbf{x}, \mathbf{x}_k) \leq r_k,\ c_k = i\} \quad \forall\, i \neq j$, where c_k is the class label of the measurement vector $\mathbf{x}_k$, and $I\{\cdot\}$ denotes the indicator function. Ties can be resolved by gradually shrinking or extending this neighborhood. Simple Euclidean metric is one very popular choice for the distance function ρ. Of course, one may use Mahalanobis distance or any other flexible and adaptive metric (see e.g., Friedman, 1994, Hastie and Tibshirani, 1996) as well. In the case of Euclidean distance, for consistency of the nearest neighbor classifier, one allows k to vary with N such that $k \to \infty$ and $k/N \to 0$ as $N \to \infty$ (see e.g., Loftsgaarden and Quesenberry 1965; Cover and Hart, 1968; Fukunaga and Hostetler, 1973; Stone, 1977; Devroye, 1981). Under the same conditions, one can show this consistency also in the case of Mahalanobis distance, when consistent estimates are used for class dispersion matrices.

Fukunaga and Hostetler (1973) and McLachlan (1992) showed the importance of finding an optimal value of k for moderately large and small sample sizes. This value is generally estimated from the training sample using cross-validation techniques. In this

chapter, instead of going for the estimation of this optimal value, we shall study the performance of different nearest neighbor classifiers indexed by k simultaneously to build up a more informative discrimination procedure. For a fixed value of k, one uses any suitable distance function to identify the neighbors and hence to determine which one of the J classes is the most favorable. We now introduce some measures for the strength of this evidence for a specific test case in favor of different competing classes and for different values of the neighborhood parameter.

4.2.1 Posterior probabilities for different populations

Given a data point $\mathbf{x}$ and a given value of the neighborhood parameter k, the proportion of observations $\sum I\{\rho(\mathbf{x}, \mathbf{x}_k) \leq r_k,\ c_k = j\}/k$ is taken as an estimate $\hat{p}_j = \hat{p}(j \mid \mathbf{x})$ for the posterior probabilities of different classes. In this chapter, to make the notations simpler, we drop the argument $\mathbf{x}$ since the dependence on $\mathbf{x}$ is obvious in all cases. For any fixed value of k, the estimated posterior probabilities determine the favorable class, and they also give an idea about the strength of discrimination. For instance, in a two class problem, larger differences in estimated posteriors indicate a high degree of strength in classification. For multi-class problems, measures like Gini's diversity index $\sum_{i \neq j} \hat{p}_i \hat{p}_j$ (see e.g., Brieman *et. al.*, 1984), entropy function $\sum_j \hat{p}_j \, log \, \hat{p}_j$ or the difference between two largest estimated posteriors can be viewed as a measure of strength. However, in a classification problem, the optimal value of k really depends on the distributions of training set observations and also on the specific observation to be classified. A single value of k often fails to give the true picture for classification of all data points. For instance if $k = 1$ is selected as the optimal value by the method of cross-validation (which may happen in practice), for each observation it will lead to posterior estimates of 0 or 1, which does not give us any idea about the strength of classification and the related uncertainties present there. Analysis using multiple values of k becomes helpful in such situations.

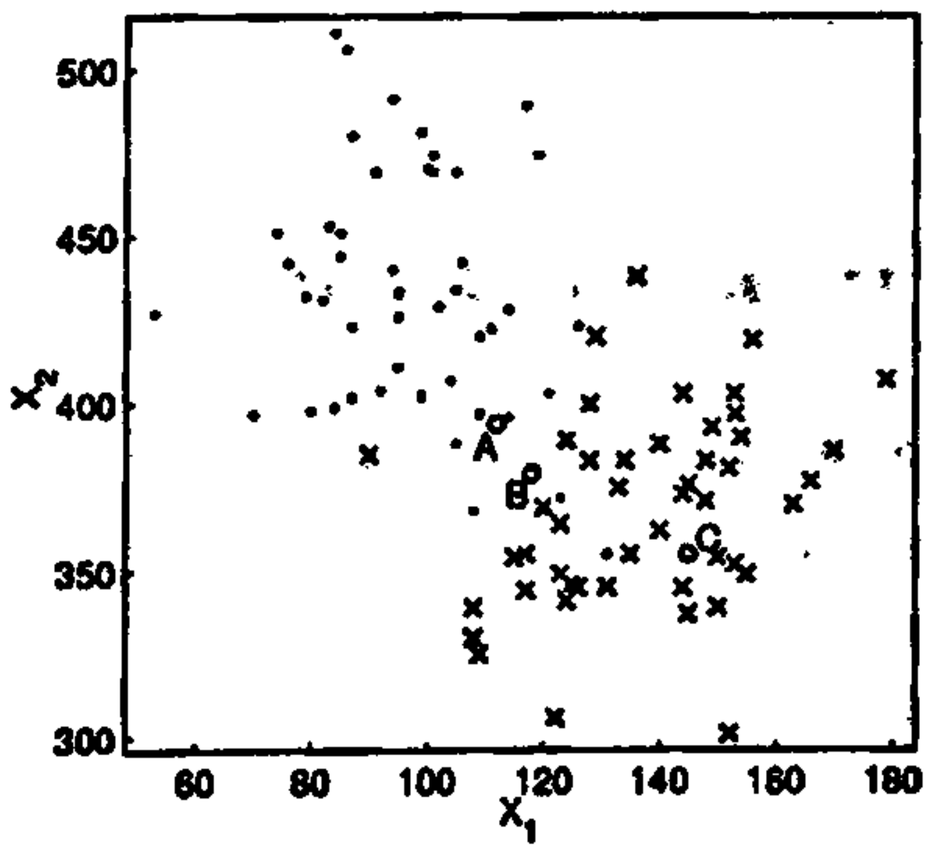


Figure 4.1 : Scatter plot for salmon data

Consider the following example of salmon data taken from Johnson and Wichern (1992). Salmon fish are born in freshwater streams, and after a year or two they swim into ocean. After a couple of years in marine water, they return to their place to spawn. 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. A scatter plot of this data set is given in Figure 4.1. We chose three observations at three different parts of the data (marked by 'o' in the figure) for which the class information is known, and we classify them using the remaining observations. Observation 'A' belongs to the Alaskan population (marked by '·') and 'C' (marked by '×') to the Canadian. One can also notice that the evidence in favor of the true class is much stronger in the second case than that in the first. Observation 'B' is taken from the second (Canadian) population but it is located near the class boundary where both the populations have almost equal strength. So, one should expect to have different behaviors of the classification methodology for these three observations. Using leave-one-out cross-validation method on this data set, we obtained $k = 7$ as the optimal neighborhood parameter, which failed to exhibit the difference in the strength of classification. It classified the observations 'A' and 'C' correctly but led to the same posterior estimates (6/7) for the true classes. Moreover, the observation 'B' got misclassified by this method. Using multiple values of k in this case, we obtained a much better result. The results of this multi-parameter analysis for these observations are given in the first column of Figure 4.2, which shows the grey scale value of posterior probabilities, where 'white' and 'black' colors represent the posterior probabilities 1 and 0, respectively. Differences in the classification results and their strength are quite evident from this figure. It clearly suggests that the strength of the evidence in favor of the true class is much higher in case of observation 'C' than that for observation 'A', which one would normally expect from Figure 4.1. Moreover, the plot for observation 'B' shows some interesting features. For small values of k, class-1 has an edge, but for higher values of k, class-2 seems to be the winner. Throughout the figure, we observe a very little difference in the posterior probability estimates of the two classes, which gives a clear indication about a border line case.

4.2.2 A Bayesian measure of strength for different populations

The idea of this measure differs critically from the usual posterior probabilities used in nearest neighbor classification. Here, instead of considering the probability of a specific class to be constant over a neighborhood around a specific point $\mathbf{x}$ and estimating it, we assume a prior distribution there to find out the posterior evidence for different classes. Suppose that $\pi(\mathbf{p})$ is the prior distribution of $\mathbf{p} = (p_1, p_2, \ldots, p_J)$ $[\sum_{j=1}^{J} p_j = 1]$. Now, for some given k, consider the k nearest neighbors of an observation $\mathbf{x}$. If t_{j_k} of these k neighbors

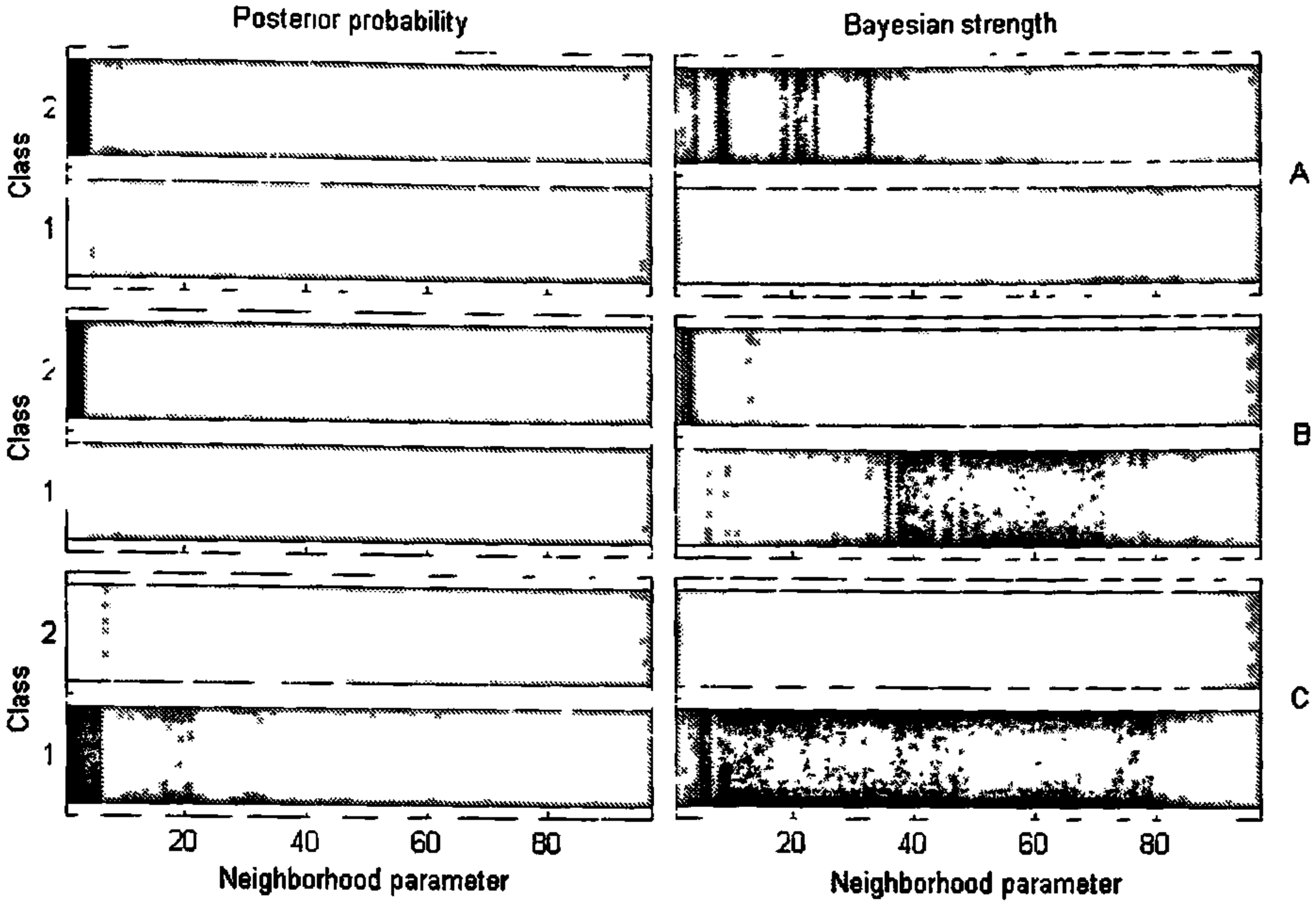


Figure 4.2 : Multi-parameter analysis of salmon data.

come from the j^{th} class, the posterior distribution of $\mathbf{t}_k = (t_{1_k}, t_{2_k}, \ldots, t_{J_k})$ $[\sum_{j=1}^{J} t_{j_k} = k]$ for given $\mathbf{p}$ and k can be expressed as

$$\varphi(\mathbf{t}_k \mid \mathbf{p}, k) = \frac{k!}{t_{1_k}!\, t_{2_k}! \ \ldots \ t_{J_k}!} \prod_{j=1}^{J} p_j^{t_{j_k}}.$$

Therefore, for some fixed k and $\mathbf{t}$, the conditional distribution of $\mathbf{p}$ is given by

$$f(\mathbf{p} \mid k, \mathbf{t}) = \pi(\mathbf{p})\ \varphi(\mathbf{t}_k \mid \mathbf{p}, k) \Big/ \int \pi(\mathbf{p})\ \varphi(\mathbf{t}_k \mid \mathbf{p}, k)\ d\mathbf{p}.$$

It is to be noted that given the value of k, $\mathbf{t}$ is well defined and therefore this conditional probability distribution depends only on k. Using this conditional distribution, we define the Bayesian measure of strength for different populations. Clearly, one will prefer the j^{th} class compared to the i^{th} one if $P\{p_j > p_i \mid k, \mathbf{t}\} > P\{p_i > p_j \mid k, \mathbf{t}\}$. Following the same idea, for a given value of k the Bayesian strength function for the j^{th} population is defined as

$$S(j \mid k) = \int_{p_j = max\ \{p_1, p_2, \ldots, p_J\}} f(\mathbf{p} \mid k, \mathbf{t})\ d\mathbf{p}.$$

Estimated posterior probabilities often fail to give a clear idea about the strength of evidence in a discrimination procedure. For instance, in a two-class problem, the posterior estimate for one class turns out to be one in all those situations where the neighbors come

from the same population, but certainly the strength is not the same in all these cases. If 10 out 10 neighbors are from the same population, that evidence should be considered as stronger than that obtained in a 1-nearest neighbor method. These differences in the strength of discrimination get well exhibited by the strength function S.

The second column of Figure 4.2 shows the grey-scale value of this Bayesian strength function for the three observations 'A', 'B' and 'C', when $\pi(\mathbf{p})$ is taken to be the probability density function of a uniform distribution on $[0, 1]$. In all these cases, this strength function leads to almost same figures as obtained in the posterior probability plots, but it enhances the evidence in favor of the winning classes. Uniform prior distribution is very easy to handle, both computationally and theoretically. Further, a uniform choice of prior is *unbiased* and *non-informative* as a prior distribution that does not distinguish among different classes. Throughout this chapter, we use the uniform prior distribution to define the Bayesian strength function. However, our empirical study suggests that the results are not much sensitive to the prior distribution and one may use many other suitable priors as well. The following theorem gives some idea about the relation between the usual and Bayesian approaches of nearest neighbor classification.

Theorem 4.1 : *For some given k, let $\hat{p}_j$ be the estimated posterior probability (as in Section 4.2.1) for the j^{th} $(j = 1, 2, \ldots, J)$ population. Also assume that $\pi(\mathbf{p})$ is symmetric in its arguments. Then, $S(j \mid k) \geq S(i \mid k)$ if and only if $\hat{p}_j \geq \hat{p}_i$.*

From this theorem, it is quite evident that the estimates of the posterior probabilities and that of the Bayesian strength functions are equivalent as far as k nearest neighbor classification is concerned, but Figure 4.2 suggests that the latter one is more preferable as a discrimination measure since it enhances the evidence in favor of the winning class. In that sense, the Bayesian strength function can be viewed as an alternative to the posterior probability which sharpens the class separability.

Strength functions have an interesting interpretation in terms of p-values as well. In the previous chapter, we have demonstrated the utility of p-value as a discrimination measure when kernel estimates of the population densities are used for classification between two competing populations. In a two-class problem, for a given value of k, except for a few tied cases, one always observes a difference in the estimated posterior probabilities of the two classes, but it is important to know whether that difference is really significant or not. The corresponding p-value gives an idea about it. If t_1 out of k neighbors of $\mathbf{x}$ come from the first class, following the idea of Chapter 3, one can use $\sum_{t<t_1} \binom{k}{t} (0.5)^t$ or $\sum_{t \leq t_1} \binom{k}{t} (0.5)^t$ as the one-sided p-value for the first population. However, it is very difficult to generalize this idea of p-values for classification problems with more than two populations. Nevertheless, we have the following result in a two-class problem.

Theorem 4.2 : *Suppose that t_1 and t_2 $(t_1 + t_2 = k)$ out of k nearest neighbors of* $\mathbf{x}$ *come form the first and the second populations, respectively. If $\pi(\mathbf{p})$ is uniform, we have*

$$\sum_{t<t_1} \binom{k}{t} (0.5)^t < S(1 \mid k) < \sum_{t \leq t_1} \binom{k}{t} (0.5)^t.$$

Therefore, the Bayesian strength function can be viewed as a p-value type measure for discrimination. When $\pi(\mathbf{p})$ is continuous, it is easy to see that $S(1 \mid k) + S(2 \mid k) = 1$, and hence $S(1 \mid k) > 0.5$ always points towards the first population.

Like the plots of p-values in Chapter 3, we have observed the figures of the strength functions to be much sharper than those of the posterior probabilities. This feature of the strength function is well explained by the following theorem.

Theorem 4.3 : *Suppose that $\pi_1, \pi_2, \ldots, \pi_J$ are the prior probabilities and f_1, f_2, ..., f_J are the density function of J classes. For a given* $\mathbf{x}$*, define $P_j = \pi_j f_j(\mathbf{x}) / \sum_{m=1}^{J} \pi_m f_m(\mathbf{x})$ as the true conditional probability of the j^{th} class $(j = 1, 2, \ldots, J)$. Now, assume that (i) $P_i > P_j$ for all $j \neq i$, and (ii) $k \to \infty$ and $k/N \to 0$ as $N \to \infty$. If $\pi(\mathbf{p})$ is symmetric in its arguments, as $N \to \infty$, $S(i \mid k) \xrightarrow{P} 1$ and $S(j \mid k) \xrightarrow{P} 0$ for all $j \neq i$.*

From the existing results (see e.g., Loftsgaarden and Quesenberry, 1965; Cover and Hart, 1968; Fukunaga and Hostetler, 1973; Stone, 1977), we know that under the conditions of Theorem 4.3, the estimates of the posterior probabilities converge to the true posteriors, which are values between 0 and 1. But in such cases, the Bayesian strength function converges either to 0 or to 1 and makes the evidence more sharper in favor of the class having the largest true posterior, and this is visible in the images in the corresponding plots.

4.2.3 Measure of uncertainty

Discrimination measures like posterior probability and Bayesian measure of strength show evidence for different classes for various choices of k. In most of the cases, from the plots of these discrimination measures the final result becomes quite transparent. However, in some situations this may not be the case. For instance, in the case of observation 'B' in salmon data, we observe strong evidence for class-1 when k is small, but for relatively larger values of k, the plot gives an indication in favor of the other class. Ranges of these two regions give a rough idea about the final classification. However, to arrive at the final decision, it is also important to know which of these two regions is more reliable. From the corresponding estimated misclassification rate, we get an idea about it. Here, we have used leave-one-out cross-validation method to estimate these misclassification rates and plotted the estimated probabilities of correct classification (re-scaled to have a minimum

value 0 and maximum value 1) for different choices of k (see Figure 4.3). This plot shows the uncertainties associated with nearest neighbor classification for different neighborhood parameters.

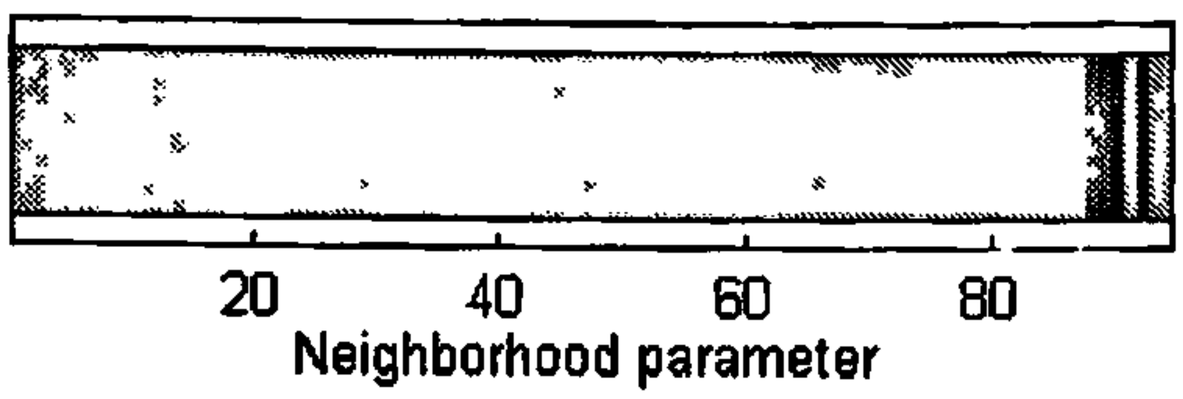


Figure 4.3 : Probabilities of correct classification (salmon data).

4.3 Aggregation of results

Information obtained at different levels of k have to be combined in a judicious way to arrive at the final decision. Bagging (see e.g., Breiman, 1996), boosting (see e.g., Friedman, Hastie and Tibshirani, 2000), arcing (see e.g., Breiman, 1998) are some of the well-known aggregation techniques that combine different classifiers to improve their performance. Breiman (1996) pointed out that nearest neighbor classifiers are more stable than neural nets (see e.g., Ripley, 1996) or classification trees (see e.g., Breiman *et. al.*, 1984), and there is not much gain in combining these classifiers using sub-sampling techniques like bagging or boosting. Shalak (1996) suggested to combine the classifiers only when they have reasonable amount of diversity among themselves. Since, two nearest neighbor classifiers with nearly same values of k are expected to produce almost similar results, it may seem that there is not much gain in combining them. However, it should also be noted that in terms of misclassification error, one would normally expect to gain by combining classifiers, and diversity in classification rules can be viewed as a measure of extent to which the error rate can be improved. Over the last few years, there is a revival of interest in aggregating nearest neighbor classifiers. Alpaydin (1997) used the condense nearest neighbor approach (see e.g., Hart, 1968) and combined a number of nearest neighbor rules developed on different representative sets from the training sample. Ho, Hull and Srihari (1994) tried to develop a multiple classifier system based on class ranks. Recently, Holmes and Adams (2002, 2003) developed a probabilistic framework for nearest neighbor classification, where they combined the nearest neighbor rules by a likelihood based method or by Bayesian analysis using Markov Chain Monte Carlo ($MCMC$) techniques (see e.g., Gilks, Richardson and Spiegelhalter, 1996). The combining procedure that we adopt in this chapter is much simpler and computationally less expensive than $MCMC$ simulations.

It is quite natural to aggregate the nearest neighbor classifiers $\{C_k;\ k = 1, 2, \ldots N\}$ taking weighted averages of the discrimination measures. The weight should be chosen in

such a way that it is higher for those values of k, which lead to lower misclassification rate $\Delta(k)$. In this chapter, we estimate $\Delta(k)$ by leave-one-out cross-validation and follow the idea of Chapter 3 to define the weight function as

$$\omega(k) = \begin{cases} e^{-\frac{1}{2}\left\{\frac{\widehat{\Delta}(k)-\Delta_0}{\sqrt{\Delta_0(1-\Delta_0)/N}}\right\}^2} & \text{if } \frac{\widehat{\Delta}(k)-\Delta_0}{\sqrt{\Delta_0(1-\Delta_0)/N}} \leq \tau \text{ and} \\ 0 & \text{otherwise,} \end{cases}$$

where N is the training sample size and $\Delta_0 = \min_k \widehat{\Delta}(k)$. Notice that, like multi-scale kernel discriminant analysis, Δ_0 and $\Delta_0(1-\Delta_0)/N$ can be viewed as estimates for the mean and the variance of the empirical misclassification rate of the best nearest neighbor classifier when it is used to classify N independent observations. The constant τ determines the maximum amount of deviation from Δ_0 in a standardized scale beyond which the weighting scheme ignores the classifier by putting zero weight on them. Clearly, $\tau = 0$ corresponds to the situation of putting all the weights only on those classifiers C_k for which $\widehat{\Delta}(k) = \Delta_0$. Because of the choice of a Gaussian-type weight function above, one does not have to consider a value of τ larger than 3 in practice. It is true that the above choice of weight function is somewhat subjective, and one may use many other suitable functions as well. However, our empirical experience suggests that the final result is not much sensitive to the weighting procedure as long as any reasonable weight function (which decreases appropriately with the cross-validated error rate) is used. Note that for a given value of k, the posterior probabilities and the Bayesian strength functions lead to the same classification rule (as indicated in Theorem 4.1). Therefore, the same weight functions can be used for averaging the posteriors and the Bayesian strength functions.

We conclude this section with the example on salmon data. The aggregated method with $\tau = 3$ correctly classified all the three observations picked earlier. Weighted average of posterior probabilities in favor of Alaskan population for these three observations were found to be 0.6261, 0.4146 and 0.1445 respectively, while we obtained 0.7777, 0.2808 and 0.0134 as weighted average of Bayesian strength for the respective cases. From these figures one can easily assess the strength of classification for these three cases.

4.4 Case studies

In this section we use some benchmark data sets to illustrate the usefulness of the proposed methodology. Some of these data sets were used in Alpaydin (1997) and Holmes and Adams (2002, 2003) for combining the nearest neighbor classification rules. We have quoted some results directly from these articles and compared the performance of our aggregation methods with them. Error rates for the best nearest neighbor rule (selected on the basis of

leave-one-out error rates in the training sample) are also reported to facilitate the comparison. Since cross-validation methods estimate the misclassification rates by naive empirical proportions, it is often possible to have more than one classifier that lead to the smallest value of $\widehat{\Delta}(k)$. In such cases, the classifier with the lowest value of k is selected. Throughout this section, we have used $\tau = 3$ for aggregating the nearest neighbor classifiers. Because of the simplicity, for our data analytic purpose, we have used simple Euclidean metric in all the cases unless otherwise specified. Most of the data sets that we use here are available at http://www.uci.ics.edu.

Salmon data : Let us start with the salmon data described in Section 4.2. This data set does not have any separate training and test sets. We partitioned it randomly to form training sets of size 30 and test sets of size 70 taking equal number of observations from the two classes. This random partition is carried out 100 times to generate 100 different training and test sets. For some of these partitions, leave-one-out method missed the optimal value of k by a large margin, and as a result it ended up with poor misclassification rates.

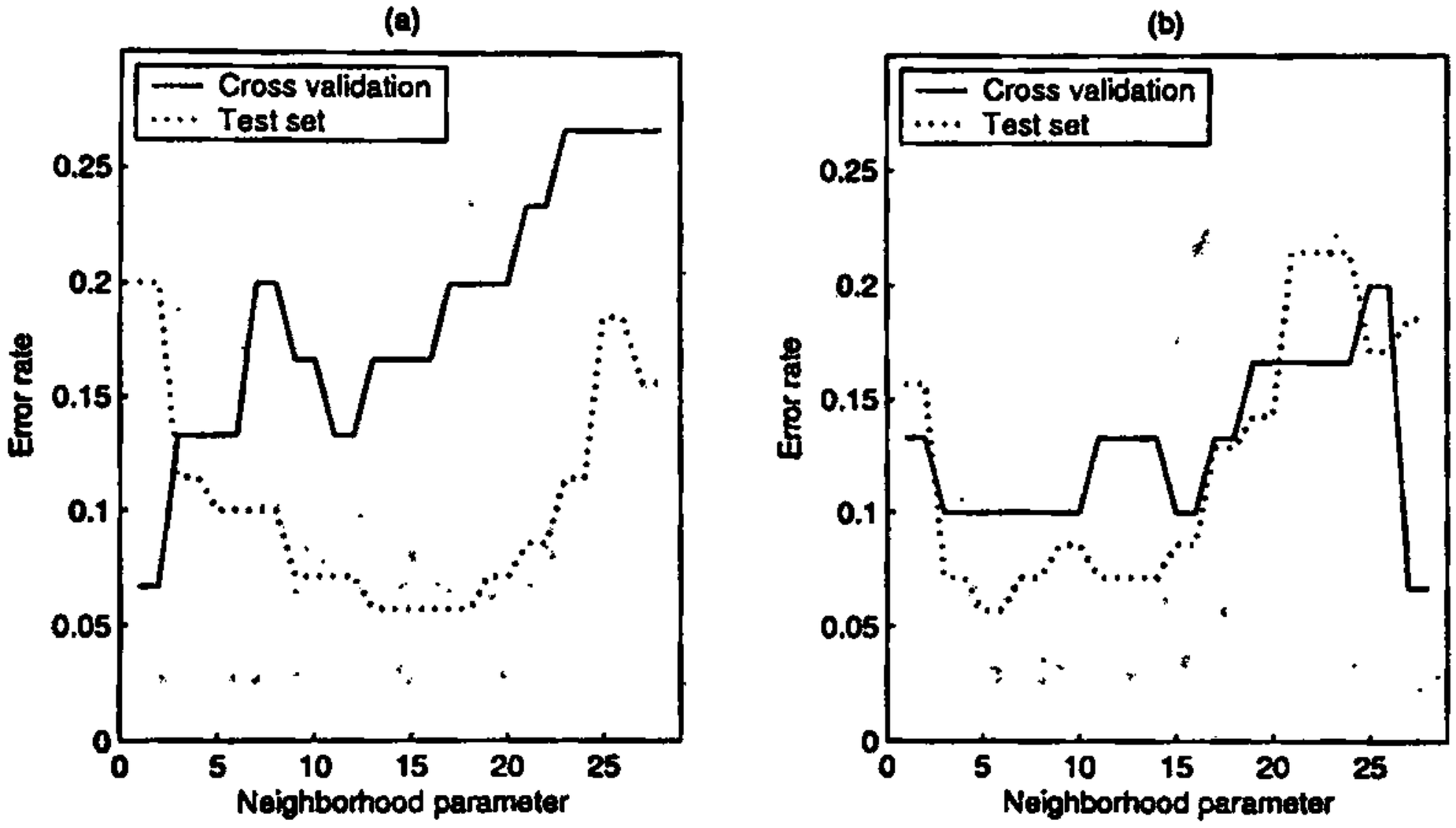


Figure 4.4 : Misclassification rates for two different partitions of salmon data.

Figure 4.4 shows the performance of leave-one-out cross-validation method for two such partitions, where this method led to test set misclassification rates of 20.00% and 18.57%, respectively. But in both the cases we could achieve significant reduction using the weighted average of posterior and weighted average of Bayesian strength function. For both these aggregating methods the test set error rates for two such typical partitions were 12.86% and 7.14% respectively. Over the 100 partitions, leave-one-out cross-validation method could achieve an average test set error rate of 10.47% with a standard error (S.E.) of 0.34%. However, for our proposed weighted average methods, we had lower misclassification rates. For weighted posterior and weighted strength aggregation methods, average test set error rates were found to be 9.39% and 9.57%, respectively, with standard errors of 0.26%

and 0.28% in the respective cases. In view of these standard errors, one can easily notice that there is a statistically significant improvement in the performance of the nearest neighbor classifier when our aggregation method is used.

Wine data : This is another example that nicely demonstrates the utility of the aggregation procedure. Description of this data set is given in Section 2.5 of Chapter 2. It contains information on 13 constituents in three different types of wines. Like salmon data, it does not have any separate test set and we carried out 100 random partitions to form 100 training and test sets consisting of 100 and 78 observations respectively. Since the measurement variables have very different scales, we used the standardized version of this data set for classification. Once again, the cross-validation method failed to find out appropriate values of k in some cases.

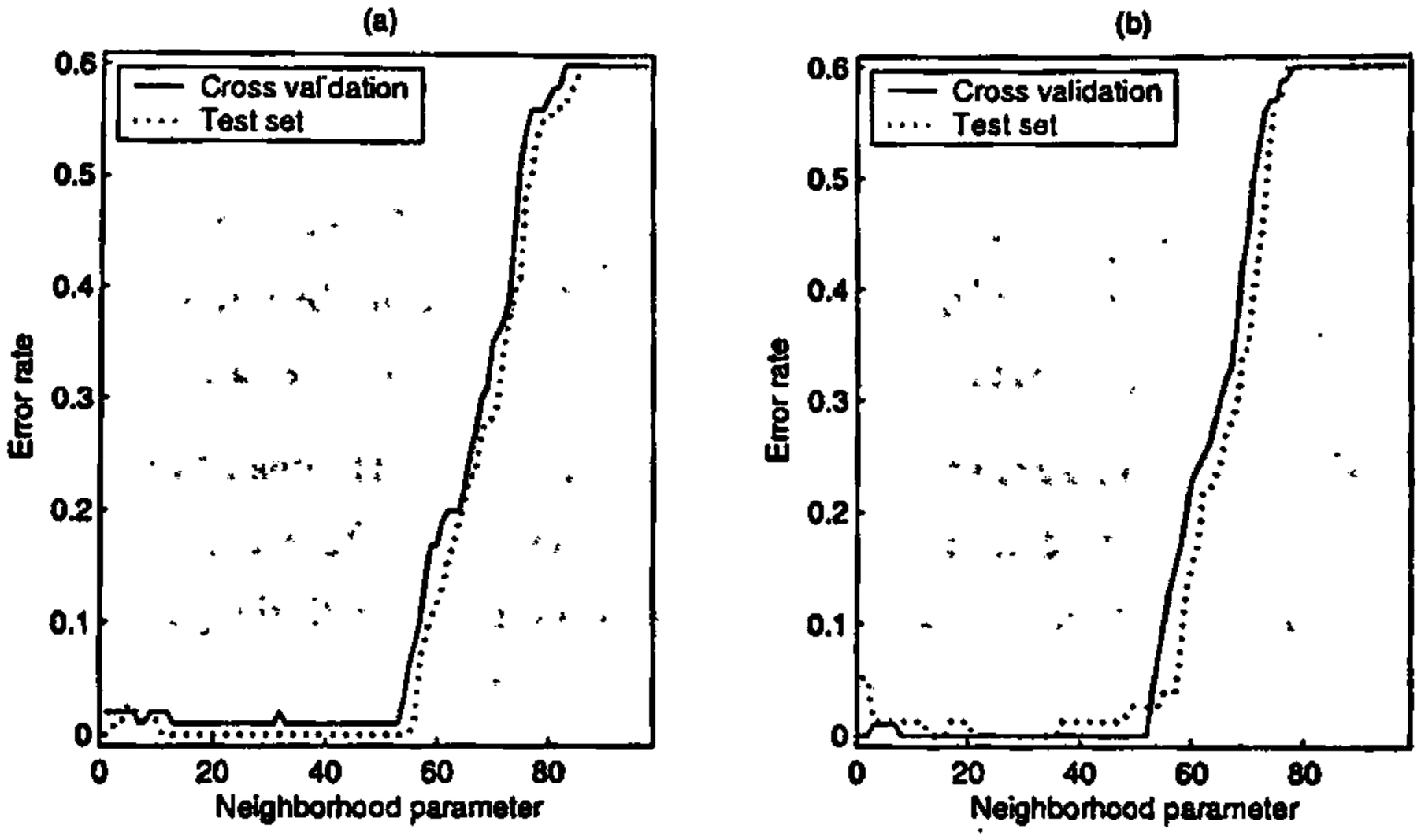


Figure 4.5 : Misclassification rates for two different partitions of wine data .

Figure 4.5 gives two such examples, where the leave-one-out cross-validation method led to error rates of 1.29% and 5.16% but in both the cases the aggregated classifiers could correctly classify all the test set observations. Since the populations are quite well separated in this data set, we observed low cross-validated errors over a wide range of k, and consequently the minimum error is obtained for a large number of choices of k. When we combine the results of only those classifiers (which is essentially an aggregation method with $\tau = 0$), even then we get a much improved performance than using a single value of k. Over the 100 partitions, leave-one-out cross-validation method led to an average test set error rate of 1.28% with a standard error (S.E.) of 0.15%. For weighted posterior and weighted strength (with $\tau = 0$), these error rates were 0.59% (with S.E. = 0.08%) and 0.38% (with S.E. =0.07%) respectively. Using $\tau = 3$, we could achieve even lower misclassification rates. Average test set error rates for $\tau = 3$ were found to be 0.53% (with S.E. = 0.08%) and 0.37% (with S.E. =0.07%) in the respective cases. The differences between the error

rates of the aggregated methods and that of the usual cross-validation procedure are clearly statistically significant.

We used a few other data sets for further comparison. Among them, vowel recognition data-1 has separate training and test sets. For each of the other two data sets (diabetes data and biomedical data), we carried out our experiments over 100 different random partitions as before and reported the average misclassification rates and their corresponding standard errors in Table 4.1. For diabetes and biomedical data, the sizes of the training samples are taken to be 50 and 100, respectively, while the rest of the observations are used to form the corresponding test sets. Description of vowel recognition data-1 and diabetes data has been given earlier in Chapter 3. Here, we give a brief description of the biomedical data.

Biomedical data set was generated by Cox and used by Cox, Johnson and Kafadar (1982) in the annual meeting on "Exposition of Statistical Graphics Technology". It contains four different measurements on each of 209 blood samples (134 for "normals" and 75 for "carriers"). 15 out of these 209 observations have some missing values. For the purpose of our analysis, we removed those 15 observations and applied the classification methods on the remaining 194 cases (127 for "normals" and 67 for "carriers").

In these data sets, the performance of the usual nearest neighbor classifier based on cross-validated estimate of neighborhood parameter and that of the proposed aggregation methods were fairly similar. Both weighted posterior and weighted strength led to almost same misclassification rates in all these cases, and their error rates were not significantly different from that of the usual nearest neighbor classifier, where the neighborhood parameter is selected by leave-one-out cross-validation.

Data sets	k-NN (cross-valid.)	Weighted posterior	Weighted strength
Salmon	10.47 (0.34)	9.39 (0.26)	9.57 (0.28)
Wine	1.28 (0.15)	0.53 (0.08)	0.37 (0.07)
Vowel-1	21.92 (2.27)	21.02 (2.23)	21.62 (2.26)
Diabetes	10.46 (0.26)	10.44 (0.29)	10.63 (0.32)
Biomedical	14.20 (0.28)	13.81 (0.25)	14.06 (0.27)

Table 4.1 : Misclassification rates (in %) for usual and combined nearest neighbor classifiers

4.4.1 Comparison with probabilistic nearest neighbor and likelihood based aggregation procedure

In this section, We have used some benchmark data sets to compare the performance of the proposed aggregation procedure with that of usual k-nearest neighbor methods (with k chosen by leave-one-out cross-validation) and other aggregation methods suggested by

Holmes and Adams (2002, 2003). Holmes and Adams (2002) developed a probabilistic framework for nearest neighbor classification, where they proposed an aggregation method based on $MCMC$ techniques. They developed another method of aggregation (see Holmes and Adams, 2003) which uses a likelihood based approach. In both these articles, the authors used some benchmark data sets to evaluate the performance of their aggregation methods. We have taken four of those data sets for comparison. Out of these four data sets, synthetic data and vowel recognition data-2 have separate training and test sets. In these cases, we reported the test set misclassification rates for different classifiers. For the other two data sets, (Pima Indian data and Australian data), the reported results are the cross-validated error rates. We partitioned these data sets into 12 and 10 folds respectively for the Pima Indian and the Australian data as it has been done in Holmes and Adams (2002, 2003). We have repeated this partitioning of the data 10 times and the average cross-validated error rates over those 10 trials are reported in Table 4.2 along with their corresponding standard errors. Note that this way of repeated partitioning leads to 120 and 100 training and test set combinations for the Pima Indian and the Australian data, respectively. In these two data sets, since the variables have very different scales, we have standardized them using an estimate of pooled dispersion matrix before classification. The result of likelihood based aggregation method on vowel data-2 was not reported in Holmes and Adams (2003), that's why we have a blank space in the table.

Since the synthetic data and the vowel data-2 have already been described in Chapter 3, a brief description of only the other two data sets is given below.

Pima Indian diabetes data : This data set consists of 768 observations coming from Pima Indian female population having at least 21 years of age. Eight measurements are taken on each individual to distinguish the diabetic patients from the normals. This data set has 268 and 500 observations from these two classes.

Australian credit data : It contains measurements on 14 different variables related to credit card applications. To protect confidentiality of the data all the variable names have been changed to meaningless symbols. This data set has 383 and 307 observations from two competing populations.

Data sets	k-NN (cross-valid.)	Likelihood (Holmes-Adams)	Prob. NN (Holmes-Adams)	Weighted posterior	Weighted strength
Synthetic	8.70 (0.89)	8.2	8.4	8.30 (0.87)	8.30 (0.87)
Vowel-2	43.72 (2.31)	49.3	*	43.72 (2.31)	43.94 (2.31)
Pima Indian	25.07 (0.36)	23.9	24.7	24.18 (0.33)	24.33 (0.32)
Australian	12.96 (0.38)	13.3	14.7	12.98 (0.35)	12.86 (0.34)

* Result is not available

Table 4.2 : Misclassification rates (in %) for different nearest neighbor classifiers

In all these data sets, the performance of the proposed aggregation methods was fairly competitive as compared to the other nearest neighbor classifiers. For the synthetic data and the Pima Indian data, our aggregation methods could achieve lower misclassification rates than that of the usual k-nearest neighbor classifier, while in the other two cases they have almost similar error rates. Both of weighted posterior and weighted Bayesian strength had nearly same misclassification rates in all these examples. The standard errors indicate that the differences between the error rates of various methods are not statistically significant.

4.4.2 Comparison with weighted CNN methods

Next, we compare our method with the performance of aggregated condensed nearest neighbor (CNN) method reported in Alpaydin (1997). Along with the vowel data-2 and the wine data, the Iris data and the glass data are also used for this comparison. Vowel data-2 has given training and test sets. For the other data sets, we used random partitioning to generate the training and test sets of the same sizes as used by Alpaydin (1997). The description of all these data sets has been given in earlier chapters of this thesis. Throughout this section, we use simple Euclidean distance for classification.

For each of these data sets, Alpaydin (1997) used CNN method on 10 representative samples taking from the training set and combined them by some weighted averaging procedure, but the author did not perform the experiment over different training and test sets. However, we divided the data sets (except for the vowel data-2, which have a given test set) randomly to form 100 different training and test sets and carried out our analysis over those 100 random partitions. Average test set misclassification rates over those 100 trials and their corresponding standard errors are reported in Table 4.3.

Data sets	k-NN (cross-valid.)	Condensed NN simple	Condensed NN weighted	NN on union	Weighted posterior	Weighted strength
Vowel-2	43.72 (2.31)	43.44	44.03	42.86	43.72 (2.31)	43.94 (2.31)
Iris	7.19 (0.35)	7.33	6.00	7.78	7.06 (0.32)	7.11 (0.30)
Wine	1.28 (0.15)	6.15	5.00	6.03	0.53 (0.08)	0.37 (0.07)
Glass	32.96 (0.49)	30.00	28.33	28.60	33.89 (0.46)	34.45 (0.54)

Table 4.3 : Misclassification rates (in %) for nearest neighbor and condensed nearest neighbor classifiers

Once again, our proposed aggregation methods showed a competitive performance in all the data sets. Apart from weighted CNN, all other methods have similar error rates for Iris data. For vowel data-2, the misclassification rates for different classifiers were also fairly equal. In the case of wine data, our proposed aggregation methods had a clear edge over the

other classifiers. Both weighted posterior and weighted strength could achieve significantly lower misclassification rates. However, these aggregation methods had a slightly higher error rates for glass data, but in view of high standard errors of the misclassification rates, their differences with the error rate of the usual nearest neighbor classifier based on cross-validated estimate of K were not statistically significant. One should also notice that in the case of glass data there are only 9, 13 and 17 observations in three competing classes, which makes it difficult to construct a good classification rule.

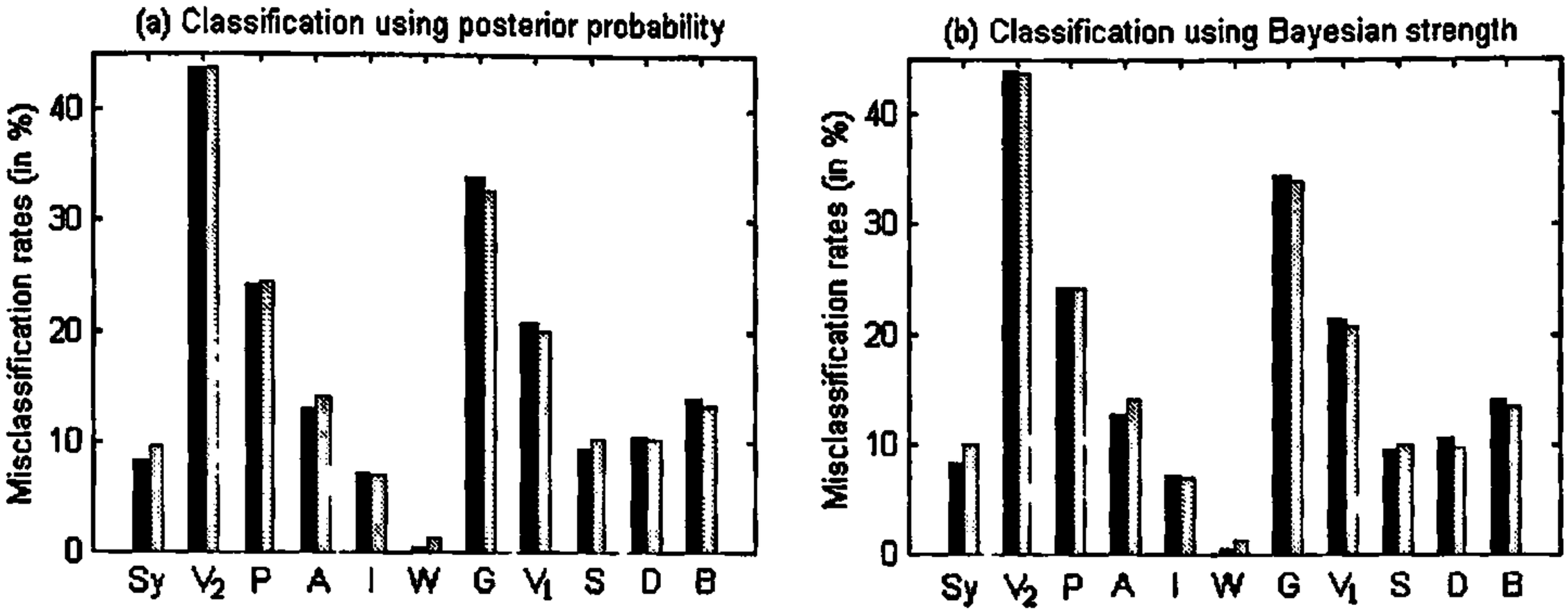


Figure 4.6 : Misclassification rates for aggregated nearest neighbor methods
Sy= Synthetic data, V_2= Vowel data-2, P= Pima Indian data, A= Australian data,
I= Iris data, W= Wine data, G= Glass data, V_1= Vowel data-1, S= Salmon data,
D= Diabetes data, B= Biomedical data

Instead of combining all nearest neighbor classifiers (k=1,2,...,N), to reduce the computational cost, one may like to restrict this set to $k \leq \sqrt{N}$. This makes the aggregation procedure much faster, and the resulting classifier in practice leads to a fairly satisfactory performance. This choice of $\sqrt{N}$ is motivated by the results on convergence of nearest neighbor estimates of posterior probabilities. It is to be noted that some other authors (see e.g., Pal, Bandopadhyay and Murthy, 1998; Mitra, Murthy and Pal, 2002) have also used the same ranges of values of the neighborhood parameter for classification purpose. Figure 4.6 gives a comparison between the performance of this truncated aggregation procedure (indicated by grey bars) and that of the original combined classifier (indicated by black bars). From this figure, it is quite evident that the truncation did not affect the misclassification rates much, but it makes an enormous savings in computation in most of the cases.

4.5 Computational aspects and related issues

This section deals with a comparison between computational complexities of different nearest neighbor classification algorithms. Since the dimension d of the measurement vector is very small as compared to the sample size N, we shall treat d to be a constant throughout this section.

Both the methods proposed by Holmes and Adams (2002, 2003) assume a logistic model for the probability functions of different classes. This assumption is somewhat subjective in nature. Our aggregation methods, on the other hand, are very straight forward, and they do not require any such model assumption. Following the idea of Holmes and Adams (see e.g., Holmes and Adams, 2002, 2003), given the value of the neighborhood parameter k and the logistic regression parameter β, one requires to perform $O(N^2)$ operations just to write down the likelihood function. Instead of using a single value of k, if one uses all possible values of k, the number of computations increases to $O(N^2 logN)$. In probabilistic nearest neighbor approach, Holmes and Adams (2002) used a prior distribution for β and k, and they adopted an $MCMC$ technique to compute the overall conditional probabilities for different competing populations. If it requires T iterations for the convergence of $MCMC$ method to get a sample (β, k), and if M such samples are required to approximate probability functions, the order of the computational complexity becomes $O(N^2 + NJTM)$. For good approximation of probability functions, one usually uses large values of M and T, which substantially increases the computational cost. On the other hand, in Holmes and Adams (2003), the authors used all possible values of k in the logistic model and adopted the iteratively re-weighted least squares method (see e.g., McCullagh and Nelder, 1989) to estimate the parameters $\beta_1, \beta_2, \ldots, \beta_N$ of the logistic function by maximizing the conditional likelihood. Since this is an iterative procedure, it is difficult to find out the exact complexity of the algorithm. However. for each iteration it requires $O(N^3)$ computations. Presence of too many parameters may make the algorithm numerically unstable as well.

Given a fixed value of k, to classify a future observation based on a training sample of size N, a k-nearest neighbor algorithm requires a computation of order $O(N)$. However, in practice, an appropriate value of k is to be estimated from the data. When leave-one-out cross-validation is used, it requires $O(N^2 logN)$ computations to find out an ideal value of k. Our proposed aggregation method based on weighted posterior requires the same order of calculation to define the weight function and to classify the future observations.

Alpaydin (1997) extracted several condensed representative sets from the training sample and used the condensed nearest neighbor (CNN) algorithm (see e.g., Hart, 1968) on each of these sets. The results are then combined using some weighting scheme to arrive at the final decision. For a given value of k, if one uses R representative sets, it requires

at least $O(kNR)$ computations to classify an observation. Depending on the size of the representative sets, the number of computations may rise up to $O(N^2R)$. However, one should notice that the formation of these representative sets depends on the ordering of the training sample points and also on the value of k. One needs to find this value of k before extracting the representative sets. Estimation of this value needs the same number of computations as that used in usual nearest neighbor method or in aggregation of weighted posteriors. CNN method involves some randomness in finding the representative sets and there is no rule to find out the required number of such sets for aggregating the classifiers for a given data set.

Our aggregation method based on weighted averaging of Bayesian strength function requires a few more computations than the weighted averaging of posteriors. Both these methods are based on the same weight functions which can be computed using $O(N^2 logN)$ calculations. But for classifying a future observation, the weighted strength procedure requires more computations than the weighted posterior method. When $J \leq 3$, it is feasible to use any numerical integration method based on appropriate averaging of the integrand over a suitable grid in the domain of integration to compute the integral appearing in the expression of the Bayesian strength function (see Section 4.2). It requires $O(NlogN + N\gamma^{J-1})$ computations to classify a future observation, where γ is the number of grid points chosen on each axis. Clearly, the computation increases rapidly with J, and therefore for $J \geq 4$, we do not adopt this procedure for computing strengths of different populations. Instead, we propose to generate a large number (M_0) of samples from the appropriate dirichlet distribution to approximate them. This method requires $O(NlogN + NM_0)$ computations for classifying a new observation into one of the J competing classes. Throughout this chapter, we have taken $M_0 = 10000$ to compute this strength function.

From, the above discussion, it is quite clear that the aggregation procedures adopted in this chapter are much simpler in nature and computationally less expensive than the other complex aggregation procedures like those based on $MCMC$ simulations or iterative re-weighted least squares. For the CNN method and for weighted averaging of posteriors, the orders of computational complexity are same as that of the usual k-nearest neighbor classification, where k is chosen by leave-one-out cross-validation. It was demonstrated in Section 4.4 that the performance of the aggregation methods is significantly better than the usual nearest neighbor method as well as CNN method on some data sets, whereas on other data sets, the difference in performance is statistically insignificant.

4.6 Classification using nearest neighbor density estimates

Instead of estimation of posteriors by naive empirical proportions, one may also find out nearest neighbor density estimates of the competing classes to develop a classification rule. This method allows the flexibility of using different values of k and different types of neighborhoods for different populations, which is quite meaningful when the populations have very different structures. If k_j is used as the neighborhood parameter for the j^{th} class, the density estimate of $f_j(\mathbf{x})$ is obtained as $\hat{f}_{j,k_j}(\mathbf{x}) = k_j/n_j V_{j,k_j}(\mathbf{x})$, where n_j is the corresponding training sample size and V_{j,k_j} is the volume of the neighborhood extending up to the k_j-th nearest neighbor of $\mathbf{x}$. Euclidean metric is the simplest choice for the distance function. One may also standardize the data set using the estimates of the class dispersion matrices and compute the density estimates for the standardized variables. Then, density estimates at the original data point can be obtained from that by using a simple transformation formula, where measurement variables undergo a linear transformation.

If one uses J different neighborhood parameters $k_1, k_2, \ldots, k_J$ for J different classes, the misclassification rate $\Delta = \Delta(k_1, k_2, \ldots, k_J)$ becomes a function of J variables which is computationally difficult to minimize. Instead, we adopt a pairwise approach, which splits the multi-class problem into several two-class problems taking a pair of classes at a time and thereby makes it computationally tractable. One can also notice that the optimal neighborhood parameter of a class density estimate not only depends on the population itself but also on its competing class densities. This pairwise approach allows the flexibility of using different values of k for a class when it is compared with different competing classes. In a two class problem, if π_1 and π_2 are their prior probabilities, and k_1 and k_2 are used as the neighborhood parameters of the two classes, $\pi_1 \hat{f}_{1,k_1}(\mathbf{x})/\{\pi_1 \hat{f}_{1,k_1}(\mathbf{x}) + \pi_2 \hat{f}_{2,k_2}(\mathbf{x})\}$ is taken as the estimate of the posterior probability for the first population. This pairwise approach also enables an effective visualization of classification results for different neighborhood parameters in a two-dimensional plot (see Figure 4.7 for example).

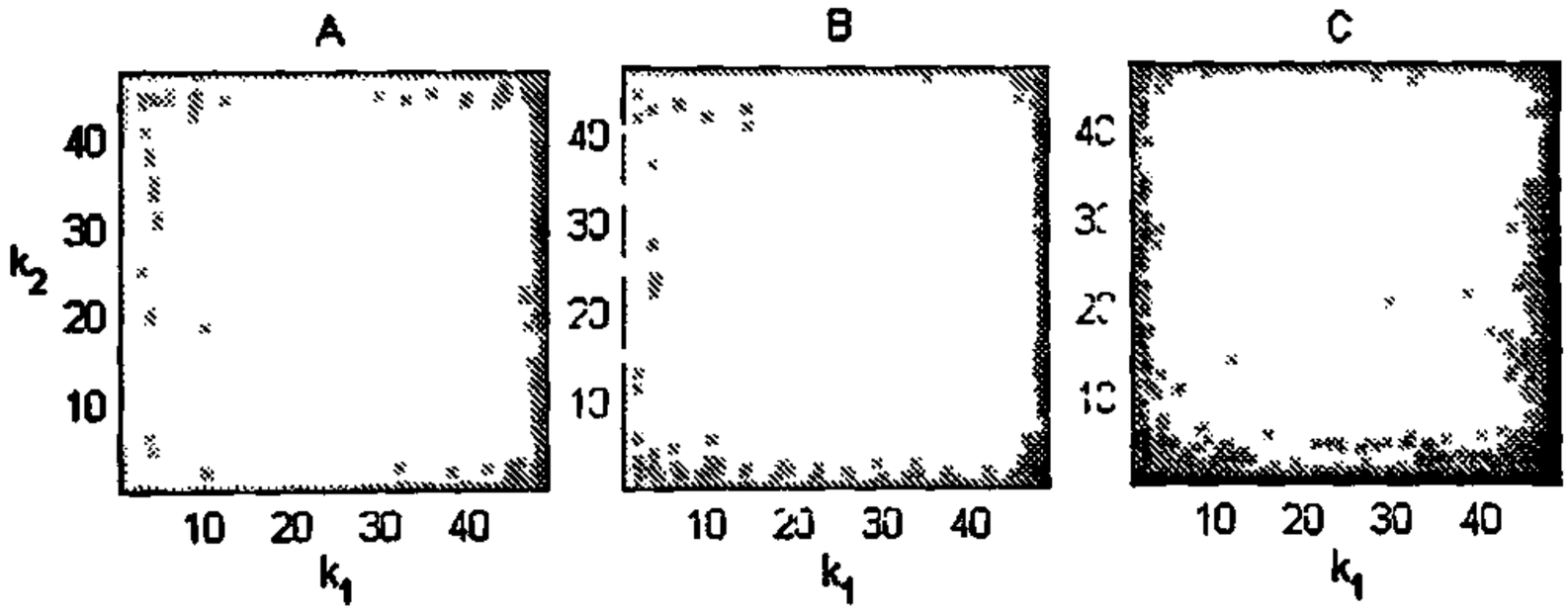


Figure 4.7 : Multi-parameter representation of posterior estimates using nearest neighbor density estimation method (salmon data)

Once again consider the salmon data, where we need to classify the observations 'A','B' and 'C' (as described in Section 4.2). Figure 4.7 gives the gray-scale representation and of the posterior probabilities (where white indicates the highest posterior and black indicates the lowest posterior) of the first population for these three observations obtained using nearest neighbor density estimates, from which the difference of the strength of classification is quite transparent. Dominance of light color in case of observation 'A' and that of dark region in case of observation 'C' in the figure makes the decision quite clear, but that is not the case for observation 'B', where we observe grey color over the entire region with a little lighter or darker shade in various parts. To find out aggregated posterior, one needs to assess the uncertainty of various regions of the plot, which is measured by the corresponding misclassification rates estimated by leave-one-out cross-validation. In Figure 4.8, we have plotted grey-scale value for the estimated probabilities of correct classification, where white color points out the regions with low misclassification rates. In order to find out the weighted posteriors, we use the same weighting scheme as used for the nearest neighbor rule (see Section 4.3), but this time we allow the pair (k_1, k_2) to have positive weight only if the corresponding misclassification rate is smaller than both the prior probabilities (i.e. the decision rule is better than a trivial classifier). This aggregation method in this salmon data yielded 0.6381, 0.3925 and 0.1030 as the weighted posteriors for the first population in the case of observations 'A','B' and 'C', respectively. This gives a clear idea about the strength of classification for these three cases. It is to be noted that for the best values of k_1, k_2 (i,e, for $k_1 = 3$ and $k_2 = 9$) that minimize the cross-validated estimate of misclassification rate, these posteriors were found to be 0.7876, 0.7180 and 0.1111 respectively. Once again the leave-one-out estimates of k_1, k_2 failed to classify the observation 'B' correctly, but the combined procedure could lead to the right decision.

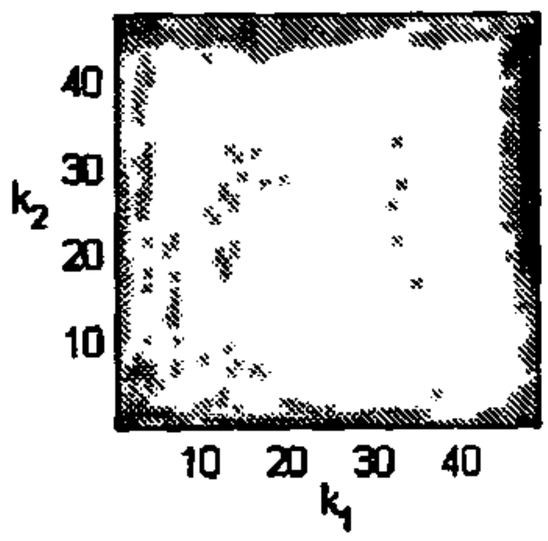


Figure 4.8 : Plot for probability of correct classification using nearest neighbor density estimates (salmon data)

For a classification problem with more than two populations, after all pairwise comparisons are done, the results have to be combined to reach the final decision. Because of its simplicity, here we use the method of majority voting (see e.g., Friedman, 1996) for the purpose of aggregation. As, we have mentioned earlier, this method may end up with tied

situation. In our subsequent discussion on the performance of the aggregated classifier, all these tied cases have been considered as misclassification. Therefore, the reported results give the error rates for the worst possible situation. Alternatively, one could use the method of pairwise coupling as discussed in Chapter 3.

We use the same data sets and the same distance functions (as used in Section 4.4) to study the performance of the aggregated classifier. Here also, each of the data sets, which do not has a separate test sample, is divided randomly to form 100 different training and test sets. The average test set error rates over those 100 partitions are reported in Table 4.4 along with their corresponding standard errors. Error rates are also given for the truncated ($k \leq \sqrt{N}$) aggregation procedure.

Once again, the combined method led to fairly competitive performance in all these data sets. In the case of vowel data-2, it could achieve an error rate of 41%, which is better than almost all the error rates reported in the existing literature. On salmon data and wine data, we could achieve a significant reduction in the error rates by aggregating the classifiers. The aggregated method produced a significantly better performance on synthetic data as well. Apart from glass data, in all other cases there was no significant difference between the error rates of the usual and the aggregated classification procedures. In most of these data sets, the truncated and computationally efficient version of the aggregation procedure could match the performance of the original untruncated aggregation procedure.

Data sets	Sample sizes		Best pair (k_1, k_2)		Weighted averaging of posteriors	
	Train	Test	N	$\sqrt{N}$	N	$\sqrt{N}$
Synthetic	250	1000	9.00(0.90)	9.20(0.91)	8.30(0.87)	9.70(0.94)
Vowel-1	338	333	20.12(2.20)	20.72(2.22)	20.12(2.20)	18.92(2.15)
Vowel-2	528	462	43.94(2.31)	43.94(2.31)	41.34(2.29)	41.34(2.29)
Pima Indian	300	468	25.54(0.15)	25.69(0.17)	24.64(0.13)	25.13(0.13)
Australian	300	390	14.97(0.13)	17.09(0.16)	15.19(0.15)	16.84(0.15)
Iris	15	135	7.16(0.33)	7.07(0.31)	7.52(0.32)	7.18(0.30)
Wine	100	78	1.75(0.14)	1.81(0.14)	0.36(0.06)	1.07(0.09)
Glass	100	114	34.30(0.39)	33.92(0.44)	36.14(0.39)	39.76(0.84)
Salmon	30	70	10.61(0.23)	10.61 (0.23)	9.71(0.24)	9.92(0.19)
Diabetes	50	95	10.54(0.25)	10.52(0.25)	10.61(0.28)	9.88(0.25)
Biomedical	100	94	14.66(0.27)	14.48(0.28)	14.04(0.25)	13.32(0.25)

Table 4.4 : Misclassification rates (in %) for classifiers based on nearest neighbor density estimates

4.7 Remarks and discussions

This chapter describes some methods for diagrammatic representation of classification results based on nearest neighbor classification rules as well as nearest neighbor density estimates. Instead of using a single value of the neighborhood parameter, here we study the results for a finite sequence of nearest neighbor methods simultaneously in order to get more useful information for classification and its strength. Diagrammatic representation enables a convenient visual comparison between the strengths of different competing populations. For nearest neighbor classification rules, where the same neighborhood parameter is used for different populations, this visual representation is quite straight forward. But, for nearest neighbor density estimation method, one has to adopt the pairwise comparison method for visual display of classification results. This pairwise approach not only reduces the computational burden, but also allows the flexibility of using different neighborhood parameters for a class when it is compared with different competing classes, and this is very desirable in practice.

The aggregation methods used here are simple and computationally less expensive than other complex aggregation procedures. As compared to the usual nearest neighbor methods, where neighborhood parameters are chosen by cross-validation techniques, these aggregation procedures produced significantly better performance on some of the data sets, while their performance on the other data sets was also quite competitive. In view of the above data analysis, it is appropriate to conclude that it would usually be better to aggregate the results of nearest neighbor methods. The gain by aggregation would be sometimes significant, and sometimes it may not be significant.

4.8 Proofs and mathematical details

Proof of Theorem 4.1 : Without loss of generality, let us assume $i < j$. From the exchangeability of p_i and p_j in $\pi(\mathbf{p})$, it is east to see that

$$\int_{p_i=max\ \{p_1,p_2,\cdots,p_J\}} \left(\prod_{m=1}^{J} p_m^{t_m}\right)\ \pi(\mathbf{p})\ d\mathbf{p} = \int_{p_j=max\ \{p_1,p_2,\cdots,p_J\}} \left(\prod_{\substack{m=1\\ m\neq i,j}}^{J} p_m^{t_m}\right) p_i^{t_j}\ p_j^{t_i}\ \pi(\mathbf{p})\ d\mathbf{p}.$$

$$\text{Now, } S(j \mid k) - S(i \mid k) = \int_{p_j=max\ \{p_1,p_2,\cdots,p_J\}} \left(\prod_{\substack{m=1\\ m\neq i,j}}^{J} p_m^{t_m}\right)\ p_i^{t_i}\ p_j^{t_i}\ (p_j^{t_j-t_i} - p_i^{t_j-t_i})\ \pi(\mathbf{p})\ d\mathbf{p}.$$

Since $p_j > p_i$ in this range, from the above expression it is quite transparent that $S(j \mid k)$ is greater (smaller) than $S(i \mid k)$ if and only if t_j is grater (smaller) than t_i.

Proof of Theorem 4.2 : When $\pi(\mathbf{p})$ is uniform, $S(1 \mid k)$ is expressed as

$$S(1 \mid k) = \int_{0.5}^{1} z^{t_1}(1-z)^{t_2}\, dz \Big/ \int_{0}^{1} z^{t_1}(1-z)^{t_2}\, dz, \quad \text{where } t_1 + t_2 = k.$$

Note that for $t_1 = 0$, the left part of the inequality, and for $t_1 = k$, the right side of the inequality gets automatically satisfied since $0 < S(1 \mid k) < 1$. For $t_1 = 1, 2, \ldots, k$ using the relation between binomial c.d.f. and incomplete beta distribution we get

$$P\{X < t_1\} = \int_{0.5}^{1} z^{t_1-1}(1-z)^{t_2} dz \Big/ \int_{0}^{1} z^{t_1-1}(1-z)^{t_2} dz. \text{ where } X \sim B(k = t_1 + t_2, 0.5)$$

Now, it is easy to see that $\int_{0.5}^{1} z^{t_1-1}(1-z)^{t_2} dz < 2\int_{0.5}^{1} z^{t_1}(1-z)^{t_2} dz$ and $\int_{0}^{0.5} z^{t_1-1}(1-z)^{t_2} dz > 2\int_{0}^{0.5} z^{t_1}(1-z)^{t_2} dz$, which imply $P\{X < t_1\} < S(1 \mid k)$.

Similarly, for $t = 0, 1, \ldots, k-1$, we have $\int_{0.5}^{1} z^{t_1}(1-z)^{t_2-1} dz > 2\int_{0.5}^{1} z^{t_1}(1-z)^{t_2} dz$ and $\int_{0}^{0.5} z^{t_1}(1-z)^{t_2-1} dz < 2\int_{0}^{0.5} z^{t_1}(1-z)^{t_2} dz$, which imply that

$$P\{X \le t_1\} = \int_{0.5}^{1} z^{t_1}(1-z)^{t_2-1} dz \Big/ \int_{0}^{1} z^{t_1}(1-z)^{t_2-1} dz > S(1 \mid k).$$

Hence, we get $\sum_{t<t_1} \binom{k}{t} (0.5)^t < S(1 \mid k) < \sum_{t \le t_1} \binom{k}{t} (0.5)^t$.

Lemma 4.1 : Suppose that $q_k(\mathbf{p}) = \prod_{m=1}^{J} p_m^{\theta_{mk}}$ ($\sum_m \theta_{mk} = 1$ for all $k = 1, 2, \ldots$) is a sequence of functions defined on $\{(p_1, p_2, \ldots, p_J) : 0 < p_1, p_2, \ldots, p_J < 1 \text{ and } \sum_m p_m = 1\}$, and $g(\mathbf{p})$ is another positive function defined on the same domain. Also assume that as $k \to \infty$, $\theta_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$ converges to $\mathbf{P} = (P_1, P_2, \ldots, P_J)$, where $\sum_m P_m = 1$ and $P_i > P_j$ for all $j \neq i$. Define a new sequence of function ζ_k on the set $C = \{(p_1, p_2, \ldots, p_j) : p_i \ge p_j \ \forall j \neq i\}$ which is given by

$$\zeta_k(\mathbf{p}) = q_k^k(\mathbf{p})\, g(\mathbf{p}) \Big/ \int_C q_k^k(\mathbf{p})\, g(\mathbf{p})\, d\mathbf{p}.$$

For every $\epsilon > 0$ and all $j \neq i$, there exist a $\delta > 0$ and a $k_0 \ge 1$ such that for all $k \ge k_0$

$$\int_{C \cap \{1-\delta \le p_j/p_i \le 1\}} \zeta_k(\mathbf{p})\, d\mathbf{p} < \epsilon.$$

Proof of Lemma 4.1 : Consider the function $q(\mathbf{p}) = \prod_{m=1}^{J} p_m^{P_m}$ defined on the same domain as that of $q_k(\mathbf{p})$. Given the values of other p_m's ($m \neq i, j$), it is easy to see that q is maximized when $p_j/p_i = P_j/P_i = 1-\eta$ $(0 < \eta < 1)$. One can also notice that q decreases with p_j/p_i when $p_j/p_i > 1-\eta$ and increases when $p_j/p_i < 1-\eta$. Now choose some $\delta < \eta/2$ and define the set $A_\delta = \{(p_1, p_2, \ldots, p_J) : 1-\delta \leq p_j/p_i \leq 1\}$. On this set q is maximized at $\mathbf{p}^\delta = (p_1^\delta, p_2^\delta, \ldots, p_J^\delta)$, where $p_m^\delta = P_m$ for all $m \neq i, j$, $p_i^\delta = (P_i + P_j)/(2-\delta)$ and $p_j^\delta = (1-\delta)(P_i + P_j)/(2-\delta)$, and this maximum value can be given by

$$\lambda(\mathbf{P}) = \Big(\prod_{\substack{m=1 \\ m \neq i,j}}^{J} P_m^{P_m} \Big) (P_i + P_j)^{P_i+P_j} (1-\delta)^{P_i} / (2-\delta)^{P_i+P_j} = M_1, \text{ say.}$$

Now, as $k \to \infty$, $\theta_{mk} \to P_m$ for all $m = 1, 2, \ldots, J$. Therefore $\theta_{jk}/\theta_{ik} \to P_j/P_i$ (since $P_i > 0$) and because of the continuity of the function λ, $|\lambda(\theta_k) - \lambda(\mathbf{P})| \to 0$. Hence, for every $\epsilon > 0$, one can always find some $k_1 \geq 1$, such that $\theta_{jk}/\theta_{ik} < 1-\delta$ and $|\lambda(\theta_k) - \lambda(\mathbf{P})| < \epsilon$ for all $k \geq k_1$. Now, note that for such values of θ_k, q_k is decreasing in p_j/p_i when $p_j/p_i \geq 1-\delta$, and on the set A_δ it has an upper bound $\lambda(\theta_k)$. Therefore,

$$\sup_{\mathbf{p} \in A_\delta \cap C} q_k(\mathbf{p}) < M_1 + \epsilon, \ \forall k \geq k_1.$$

On the other hand, the function q is maximized at $\mathbf{p}^0 = \mathbf{P}$, which is an interior point of C. Let us denote this value $\prod_{m=1}^{J} P_m^{P_m}$ by M_2. Clearly, $M_2 > M_1$ and because of continuity, for every $\epsilon > 0$ it is possible to find ϵ_0 such that $\|\mathbf{p} - \mathbf{P}\| + \|\theta_k - \mathbf{P}\| < \epsilon_0 \Rightarrow |\prod_{m=1}^{J} p_m^{\theta_{mk}} - \prod_{m=1}^{J} P_m^{P_m}| < \epsilon$, where $\| \ \|$ denotes the usual Euclidean norm. Since θ_k converges to $\mathbf{P}$, it is always possible to choose a ball $B \subset C$ of radius $\epsilon_1 < \epsilon_0$ around $\mathbf{P}$ such that for some $k_2 \geq 1$ and all $k > k_2$,

$$\inf_{\mathbf{p} \in B \cap C} q_k(\mathbf{p}) > M_2 - \epsilon \ \forall k \geq k_2.$$

Note that the above results hold for every epsilon $\epsilon > 0$. Choose an ϵ such that $(M_1 + \epsilon)/(M_2 - \epsilon) < t$ for some $t < 1$. Define

$$\alpha = \int_{A_\delta \cap C} g(\mathbf{p}) \, d\mathbf{p} \quad \text{and} \quad \beta = \int_{B \cap C} g(\mathbf{p}) \, d\mathbf{p}.$$

Now it is quite easy to see that for all $k > k^* = max\{k_1, k_2\}$,

$$\int_{A_\delta \cap C} \zeta_k(\mathbf{p}) \, d\mathbf{p} < \int_{A_\delta \cap C} q_k^k(\mathbf{p}) \, g(\mathbf{p}) \Big/ \int_{C} q_k^k(\mathbf{p}) \, g(\mathbf{p}) \, d\mathbf{p} \ < \ \alpha t^k / \beta.$$

To arrive at the final result choose $k_0 > k^*$ such that $\alpha t^k/\beta < \epsilon$.

Proof of Theorem 4.3 : Take $\theta_{mk} = t_{m_k}/k$ for $m = 1, 2, \ldots, J$, $g(\mathbf{p}) = \pi(\mathbf{p})$ and consider the sequence of functions q_k and ζ_k as described in Lemma 4.1. From the existing results

(see e.g., Loftsgaarden and Quesenberry, 1965; Cover and Hart, 1968; Stone, 1977; Devroye, 1981), we know that if $k \to \infty$ and $k/N \to 0$ as $N \to \infty$, then θ_k converges in probability to $\mathbf{P}$, the vector of true conditional probabilities (at $\mathbf{x}$) of different classes. Now, following the idea of exchangeability of p_i and p_j as used in the proof of Theorem 4.1, it is easy to see that

$$S(j \mid k)/S(i \mid k) = \int_C (p_j/p_i)^{k(\theta_{ik}-\theta_{jk})} \zeta_k(\mathbf{p})\, d\mathbf{p},$$

where ζ_k and C have the same meaning as in Lemma 4.1. For $0 < \delta < 1$, define $A_\delta = \{(p_1, p_2, \ldots, p_J) : 1 - \delta \leq p_j/p_i \leq 1\}$. Since $p_j/p_i \leq 1$ on C and $k(\theta_{ik} - \theta_{jk}) \xrightarrow{P} \infty$ as $k \to \infty$, for every $\epsilon, \lambda > 0$, it is possible to find $\delta > 0$ and $k_0 \geq 1$ (see Lemma 4.1) such that

$$P\left\{ \int_{A_\delta \cap C} (p_j/p_i)^{k(\theta_{ik}-\theta_{jk})}\, \zeta_k(\mathbf{p})\, d\mathbf{p} < \int_{A_\delta \cap C} \zeta_k(\mathbf{p})\, d\mathbf{p} < \epsilon/2 \right\} > 1 - \lambda/2 \ \ \forall\, k \geq k_0.$$

Again, note that on the set $A_\delta^c \cap C$, $(p_j/p_i)^{k(\theta_{ik}-\theta_{jk})}$ uniformly converges to 0 in probability. Therefore, for those same ϵ and λ, one can find some $k_1 \geq 1$ such that

$$P\left\{ \sup_{A_\delta^c \cap C} (p_j/p_i)^{k(\theta_{ik}-\theta_{jk})} < \epsilon/2 \right\} > 1 - \lambda/2 \ \ \forall\, k \geq k_1.$$

Hence, $P\{S(j \mid k)/S(i \mid k) < \epsilon\} > 1 - \lambda$ for all $k > max\{k_0, k_1\}$. This implies $S(j \mid k) \xrightarrow{P} 0$ for all $j \neq i$ (since $S(i \mid k) < 1$). Now, the result $S(i \mid k) \xrightarrow{P} 1$ follows from the facts that $\sum_{j=1}^{J} S(j \mid k) = 1$ and J is finite.

Chapter 5

Data depth and discriminant analysis using finite dimensional parametric surfaces

5.1 Main problem and motivation

Any discriminant analysis problem can be viewed as a problem of finding an appropriate function $Q(\mathbf{x})$ of the measurement variables $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ that contains the maximum information about class separability. In a two-class problem, a function Q can be used to construct the discriminating surface $S = \{\mathbf{x} : Q(\mathbf{x}) = 0\}$ between the two classes. For instance, in linear discriminant analysis, one tries to determine a separating hyperplane $S = \{\mathbf{x} : \alpha'\mathbf{x}+\beta = 0\}$ based on the training sample. Several methods for choosing the projection vector α and the constant β from the training sample are available in the literature (see e.g., Fukunaga, 1990; McLachlan, 1992; Duda *et. al.*, 2000; Hastie *et. al.*, 2001). Similarly, in quadratic discriminant analysis, one uses a quadratic separating surface $S = \{\mathbf{x} : \alpha'\mathbf{x} + \mathbf{x}'\Lambda\mathbf{x} + \beta = 0\}$, where Λ is a symmetric matrix to be chosen from the training sample in addition to α and β. Fisher's original approach in linear and quadratic discriminant analysis (see Fisher, 1936) was primarily motivated by multivariate normal distribution of the measurement vector $\mathbf{x}$, and there estimates for α , β and Λ were constructed using the mean vectors and the dispersion matrices of the training sample observations. Under the assumption of multivariate normal distribution for the data, such linear or quadratic classifiers turn out to be the optimal Bayes classifiers. Further, such classifiers based on simple parametric surfaces are useful in the sense that they lead to a good low dimensional view of class separability. However, since such methods require the estimation of α, β and Λ using the first and the second order moments of the training samples, these procedures are

not very robust and happen to be highly sensitive to extreme values and outliers if they are present in the training sample. When the assumption of normal distribution for the data is violated, these methods may lead to a rather poor classification especially if the observations follow some distribution having heavy tails. On the other hand, nonparametric methods discussed in the preceding chapters do not use any information on the data distribution and the form of the separating surface. These flexible methods of discriminant analysis, however, sometimes have a tendency to overfit the training data, and their performance on test cases may not be very good unless a very large training data set is available.

In this chapter, we will study some semiparametric classification methods that uses the notions of half-space depth (Tukey, 1975) and regression depth (Rousseeuw and Hubert, 1999) functions to construct linear and nonlinear separating surfaces among the competing populations. Over the last few years, various notions of data depth have emerged as powerful exploratory and inferential tools for nonparametric multivariate analysis (see e.g., Chaudhuri and Sengupta, 1993; Liu, Parelius and Singh, 1999; Vardi and Zhang, 2000; Zuo and Serfling, 2000a, 2000b; Serfling, 2002). Recently, Christmann and Rousseeuw (2001) and Christmann, Fischer and Joachims (2002) used regression depth for constructing linear classifiers in two-class problems and investigated their statistical performance. They also made some comparative studies of such linear classifiers with the classifiers built using support vector machines (see e.g., Vapnik, 1995, 1998). Since the discriminant analysis tools investigated in this chapter are based on half space depth and regression depth functions, they are completely distribution free in nature. These classifiers use the distributional geometry of the multivariate data cloud formed by the training sample to minimize the empirical misclassification rates, and they are not dependent on any specific model for the underlying population distributions.

5.2 Description of the methodology

Half space depth of a point in the multi-dimensional space measures the centrality of that point with respect to a multivariate distribution or a given multivariate data cloud. Regression depth, on the other hand, is a concept of depth of a regression fit (i.e., a line or a hyperplane). Hyperplanes are the simplest forms of separating surfaces, which lead to linear discrimination among the classes. We now describe how these two different depth based linear classification tools are built using a given training sample with two classes. Subsequently, we will generalize these techniques to nonlinear classification as well as to multi-class discrimination problems.

5.2.1 Linear classification using half-space depth

Half-space depth (see e.g., Tukey, 1975; Donoho and Gasko, 1992) of a d-dimensional observation $\mathbf{x}$ with respect to a multivariate distribution F is defined as the minimum probability of a closed half-space containing $\mathbf{x}$.

$$HD(\mathbf{x}, F) = \inf_{H} P_F\{H : H \text{ is a closed half-space and } \mathbf{x} \in H\}$$

Empirical version of this depth function is obtained by replacing F with the empirical distribution function F_n. Half-space depth is affine invariant, and its empirical version uniformly converges to the population depth function when F is continuous. Different properties of this depth function have been studied extensively in the literature (see e.g, Nolan, 1992; Donoho and Gasko, 1992; He and Wang, 1997; Zuo and Serfling, 2000a, 2000b).

Suppose that we have a two-class problem with univariate data. If the classes are well-separated, we would expect that most of the observed differences $x_{1i} - x_{2j}$ (x_{1i} and x_{2j} belong to two different classes for $1 \leq i \leq n_1$, $1 \leq j \leq n_2$) will have the same sign (positive or negative). This idea can be easily extended to multivariate situations, where if the two classes can be well discriminated by a linear discriminant function, we would expect to have a linear projection $\boldsymbol{\alpha}'\mathbf{x}$ for which most of the differences $\boldsymbol{\alpha}'\mathbf{x}_{1i} - \boldsymbol{\alpha}'\mathbf{x}_{2j}$ will have the same sign. We propose to estimate $\boldsymbol{\alpha}$ by maximizing

$$U_{\mathbf{n}}(\boldsymbol{\alpha}) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\{\boldsymbol{\alpha}'(\mathbf{x}_{1i} - \mathbf{x}_{2j}) > 0\},$$

where $\mathbf{n} = (n_1, n_2)$ is the vector of sample sizes for the two classes. Clearly, the maximization problem can be restricted on the set $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| = 1\}$. It can also be shown that this is actually a maximization problem over a finite set (see e.g., Chaudhuri and Sengupta, 1993), and the estimated linear projection is orthogonal to the hyperplane, which defines the half-space depth of the origin with respect to the data cloud formed by the differences $\mathbf{x}_{1i} - \mathbf{x}_{2j}$ in the d-dimensional space. This generalized U-statistic $U_{\mathbf{n}}(\boldsymbol{\alpha})$ is a measure of linear separability between the two classes along the direction $\boldsymbol{\alpha}$, and its maximum value over different possible choices of $\boldsymbol{\alpha}$ can be viewed as a multivariate analog of the well known univariate Mann-Whitney U-statistic (or Wilcoxon's two sample rank statistic). The maximizer of $U_{\mathbf{n}}(\boldsymbol{\alpha})$, denoted by $\hat{\boldsymbol{\alpha}}_H$, can be used to construct a linear classifier of the form $\hat{\boldsymbol{\alpha}}_H'\mathbf{x} + \beta = 0$ for some suitably chosen constant β. The classification rule, and consequently the corresponding misclassification probabilities depend on the choice of this constant. After getting the estimate $\hat{\boldsymbol{\alpha}}_H$, $\hat{\beta}_H$ can be obtained by minimizing w.r.t. β the average training set misclassification error $\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H, \beta)$ given by the expression

$$\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H, \beta) = \frac{\pi_1}{n_1} \sum_{i=1}^{n_1} I\{\hat{\boldsymbol{\alpha}}_H'\mathbf{x}_{1i} + \beta < 0\} + \frac{\pi_2}{n_2} \sum_{i=1}^{n_2} I\{\hat{\boldsymbol{\alpha}}_H'\mathbf{x}_{2i} + \beta > 0\},$$

where π_1 and π_2 are the prior probabilities for the two classes.

5.2.2 Linear classification using regression depth

Regression depth (see e.g., Rousseeuw and Hubert, 1999; Bai and He, 1999) gives the depth of a 'fit' determined by a vector $\alpha_+ = (\alpha, \beta) \in R^{d+1}$ of co-efficients in a linear regression framework. Given a data cloud $\Omega_n = [\{\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id}), y_i\};\ i = 1, 2, \ldots, n]$ of n observations, α_+ is called a 'nonfit' to Ω_n if and only if there exists an affine hyperplane H in the x-space such that no $\mathbf{x}_i$ belongs to H, and the residuals $y_i - \alpha_+'(\mathbf{x}_i, 1)$ are all positive in one open half space (i.e., one side of H) in the x-space and all negative in the complementary open half space (i.e., the other side of H). Regression depth of a 'fit' α_+ is defined as the minimum number of observations that need to be removed to make it a 'nonfit'.

Recently, Christmann *et. al.* (2002) used this idea of regression depth in a binary regression context to construct linear classifiers for two-class problems. If we take the class-labels ('0' or '1') as the values of the response variable y, and consider a 'fit' $\alpha_+ = (0, 0, \ldots, 0, 0.5)$, α_+ will be a nonfit to Ω_n if and only if there exists a hyperplane H in the x-space, which completely separates the data points from the two classes. Hence, the regression depth of the 'fit' α_+ can be viewed as the minimum number of misclassifications that can be achieved by a separating hyperplane H in the x-space.

Since Christmann *et. al.* (2002) considered only the problem of determining the separating hyperplane by minimizing the total count of misclassified observations, their linear classifier is empirically optimal when the two competing classes have prior probabilities proportional to their training sample sizes. In the general case, one can properly adjust the weights for the different observations and define the weighted regression depth of a 'fit' α_+ as the minimum amount of weights that need to be removed to make it a 'nonfit'. Then the weighted regression depth eventually turns out to be the average training sample misclassification probability

$$\Delta_{\mathbf{n}}(\alpha, \beta) = \frac{\pi_1}{n_1} \sum_{i=1}^{n_1} I\{\alpha' \mathbf{x}_{1i} + \beta < 0\} + \frac{\pi_2}{n_2} \sum_{i=1}^{n_2} I\{\alpha' \mathbf{x}_{2i} + \beta > 0\}\,.$$

Here, the minimization of $\Delta_{\mathbf{n}}(\alpha, \beta)$ w.r.t. α and β gives the estimates $\hat{\alpha}_R$ and $\hat{\beta}_R$ defining the separating hyperplane to be used for classification. Once again, it is clear that the minimization problem can be restricted to $\{(\alpha, \beta) : \|(\alpha, \beta)\| = 1\}$. It is also straight forward to verify that the minimization of $\Delta_{\mathbf{n}}(\alpha, \beta)$ actually turns out to be an optimization problem over a finite set (see e.g. Rousseeuw and Struyf, 1998).

Christmann *et. al.* (2002) discussed the fact that the maximum likelihood estimate in a logistic regression problem exists only when there is some overlap in the covariate space (the x-space) between the data points from the two classes corresponding to the values 0 and 1 of the response variable (see e.g., Albert and Anderson, 1984; Santner and Duffy,

1986). In completely separable cases, there exists no finite maximum likelihood estimate for regression coefficient vector α_+. If the observations from the two classes are completely separable, it is fairly easy to see that $(\hat{\alpha}_R, \hat{\beta}_R)$ is a minimizer of $\Delta_{\mathbf{n}}(\alpha, \beta)$ if and only if $\hat{\alpha}_R$ maximizes $U_{\mathbf{n}}(\alpha)$, and hence this $\hat{\alpha}_R$ is also an $\hat{\alpha}_H$.

5.2.3 Depth based classification using nonlinear surfaces

In practice, linear classifiers may be inadequate when the class boundaries are more complex in nature. In such situations, one has to depend on nonlinear separating surfaces for discriminating among the competing classes. To construct such surfaces, we can project the observations $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ into a higher dimensional space to have the new vector of measurement variables $\mathbf{z}_i = (\vartheta_1(\mathbf{x}_i), \vartheta_2(\mathbf{x}_i), \ldots, \vartheta_r(\mathbf{x}_i))$, and perform a linear classification on that r-dimensional space. For instance, if we project the observations to the space of quadratic functions, it can be viewed as a linear classification with $r = d + \binom{d}{2}$ measurement variables, which eventually give rise to a quadratic separation in the original d-dimensional space. The quantities

$$U_{\mathbf{n}}(\alpha) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\{\alpha'(\mathbf{z}_{1i} - \mathbf{z}_{2j}) > 0\} \text{ and}$$

$$\Delta_{\mathbf{n}}(\alpha, \beta) = \frac{\pi_1}{n_1} \sum_{i=1}^{n_1} I\{\alpha' \mathbf{z}_{1i} + \beta < 0\} + \frac{\pi_2}{n_2} \sum_{i=1}^{n_2} I\{\alpha' \mathbf{z}_{2i} + \beta > 0\}$$

can be optimized as before to get appropriate estimates of α and β, which are to be used to form the discriminating surface in a two-class problem.

As we have already mentioned, traditional methods of linear and quadratic discriminant analysis are primarily motivated by multivariate normal distributions. As a matter of fact, in a two-population problem, the moment based linear discriminant function is closely related to the Hotelling's T^2 or Mahalanobis distance, which are well known to be sensitive to possible outliers present in the data. On the other hand, the distribution free depth based classifiers discussed above are quite robust against such outliers, and we will now illustrate this using a small example. We consider a binary classification problem, where both the population distributions are bivariate normal with mean vectors $\mu_1 = (0, 0)$ and $\mu_2 = (2, 2)$, and they have a common dispersion matrix $\Sigma = \mathbf{I}_2$. A random sample of size 50 is generated from each of the classes to form the training sample. As the optimal Bayes rule is linear for this problem, a good linear classifier is expected to give a good separation of the data from the two populations. Here the traditional (shown as *LDA*) and the two depth based linear classifiers (shown as H-depth and R-depth) performed quite well in discriminating between the two populations (see Figure 5.1(a)). But the scenario gets completely changed when five of the class-1 observations get replaced by outliers generated

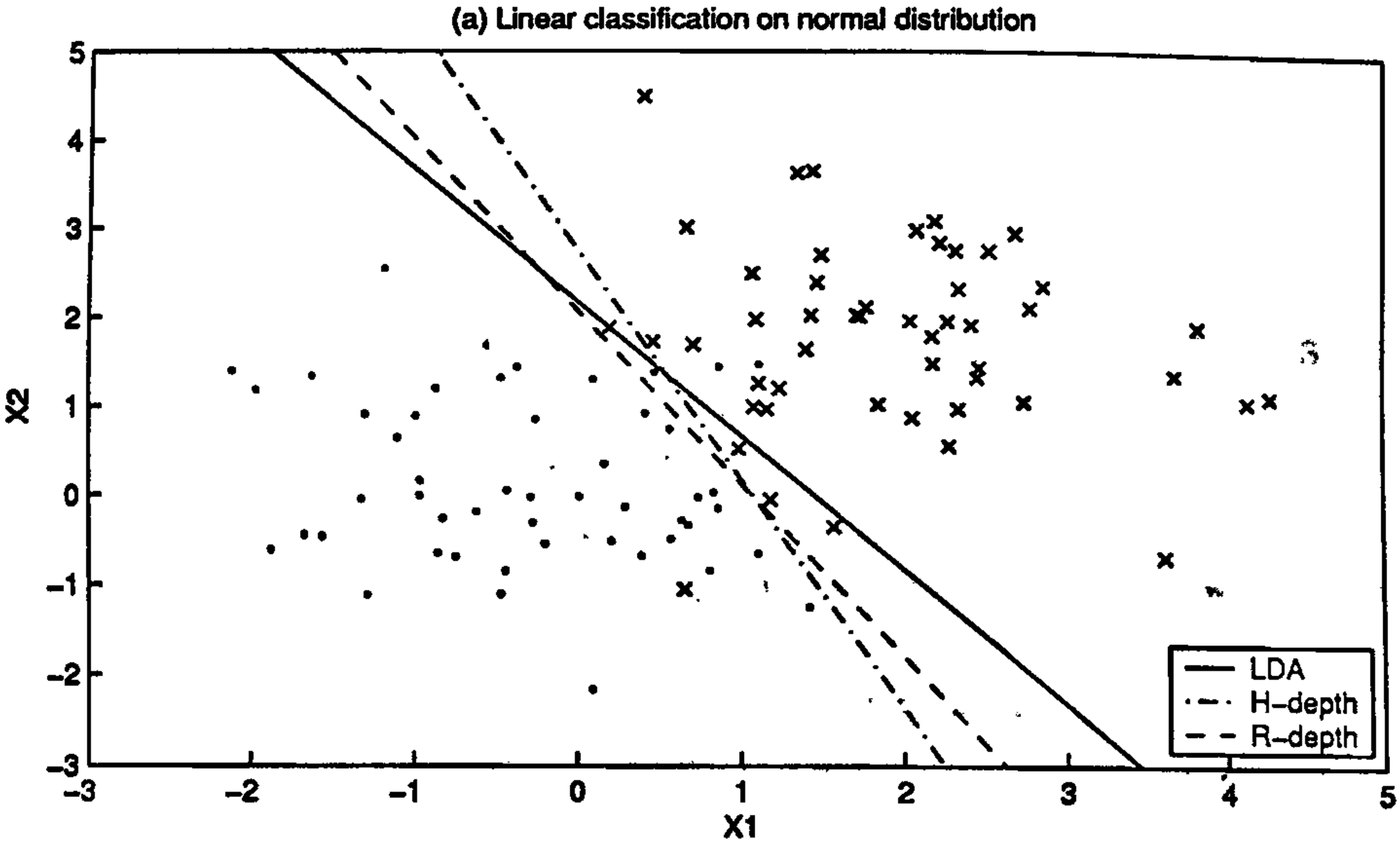


Figure 5.1 : Different linear classifiers for normal and perturbed normal distributions

from $N_2(10, 10, 1, 1, 0)$. In the presence of this contamination, the performance of the traditional moment based linear discriminant function falls drastically (see Figure 5.1(b)), but the two depth based distribution free classifiers remain more or less unaffected. For such a bivariate example, the outliers are clearly visible in the scatter-plot, but for multivariate data in higher dimensions that may not be the case. So, it is important to have classifiers that have some automatic safeguards against such outliers which may or may not be easily identified using any available diagnostic tool.

5.3 Large sample properties of depth based classifiers

We will now discuss the asymptotic behavior of the classifiers based on half-space and regression depths as the size of training sample grows to infinity. As before, suppose that we have a two class problem, and $\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$ and $\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$ are two independent sets of d-dimensional i.i.d. observations from two d-dimensional competing populations. Let $\mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{1n_1}$ and $\mathbf{z}_{21}, \mathbf{z}_{22}, \ldots, \mathbf{z}_{2n_2}$ be their transformations into the r-dimensional space as described in the previous section, and $\hat{\boldsymbol{\alpha}}_H$ is a maximizer of $U_{\mathbf{n}}(\boldsymbol{\alpha})$ while $\hat{\boldsymbol{\alpha}}_R$, $\hat{\beta}_R$ are minimizers of $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ as before.

Theorem 5.1 : *Assume that as* $N = n_1 + n_2 \to \infty$, $n_1/N \to \lambda$ $(0 < \lambda < 1)$. *Define,* $U(\boldsymbol{\alpha}) = Pr\{\boldsymbol{\alpha}'(\mathbf{z}_{1i} - \mathbf{z}_{2j}) > 0\}$ *and* $\Delta(\boldsymbol{\alpha}, \beta) = \pi_1 Pr\{\boldsymbol{\alpha}'\mathbf{z}_{1i} + \beta < 0\} + \pi_2 Pr\{\boldsymbol{\alpha}'\mathbf{z}_{2j} + \beta > 0\}$. *Then, as* $N \to \infty$, *we have*

$(i) |U_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H) - \max_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha})| \xrightarrow{a.s.} 0$ *as well as* $|U(\hat{\boldsymbol{\alpha}}_H) - \max_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha})| \xrightarrow{a.s.} 0$, *and*

$(ii) |\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_R, \hat{\beta}_R) - \min_{\boldsymbol{\alpha}, \beta} \Delta(\boldsymbol{\alpha}, \beta)| \xrightarrow{a.s.} 0$ *as well as* $|\Delta(\hat{\boldsymbol{\alpha}}_R, \hat{\beta}_R) - \min_{\boldsymbol{\alpha}, \beta} \Delta(\boldsymbol{\alpha}, \beta)| \xrightarrow{a.s.} 0$.

Further, when there exist unique optimizers $\boldsymbol{\alpha}^*_H$ *and* $(\boldsymbol{\alpha}^*_R, \beta^*_R)$ *for* $U(\boldsymbol{\alpha})$ *and* $\Delta(\boldsymbol{\alpha}, \beta)$ *respectively, and* U *and* Δ *are continuous functions of their arguments,* $\hat{\boldsymbol{\alpha}}_H$ *converges to* $\boldsymbol{\alpha}^*_H$ *and* $(\hat{\boldsymbol{\alpha}}_R, \hat{\beta}_R)$ *converges to* $(\boldsymbol{\alpha}^*_R, \beta^*_R)$ *almost surely as* $N \to \infty$.

Here, $U(\boldsymbol{\alpha})$ is a measure of linear/non-linear separability between two competing multivariate distributions along the direction $\boldsymbol{\alpha}$, and $\max_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha})$ measures the maximum linear/non-linear separability between two multivariate populations. Note also that $\Delta(\boldsymbol{\alpha}, \beta)$ is the average misclassification probability when the surface $\boldsymbol{\alpha}'\mathbf{z} + \beta = 0$ is used to discriminate between the two competing populations, and $\min_{\boldsymbol{\alpha}, \beta} \Delta(\boldsymbol{\alpha}, \beta)$ is the best average misclassification probability achievable using such linear/non-linear classifiers. It will be appropriate to point out here that $\Delta(\hat{\boldsymbol{\alpha}}_R, \hat{\beta}_R)$ can be viewed as the conditional average misclassification probability given the training sample, when the surface $\hat{\boldsymbol{\alpha}}'_R\mathbf{z} + \hat{\beta}_R = 0$ is used to classify a future observation coming from one of the two competing populations. A proof of this

theorem will be given in the Section 5.8. We state below some interesting and useful results for depth based linear and nonlinear classifiers that follow from this theorem.

Corollary 5.1 : *The average misclassification probability of the regression depth based linear (or nonlinear) classifier asymptotically converges to the best possible average misclassification rate that can be obtained using a linear (or nonlinear) classifier as the training sample size tends to infinity. Further, when the best linear (or nonlinear) classifier is unique, the regression depth based linear (or nonlinear) classifier itself converges to that optimal discriminating hyperplane (or nonlinear surface) almost surely.*

Corollary 5.2 : *Suppose that the population densities f_1 and f_2 of the two competing classes are elliptically symmetric with a common scatter matrix Σ. Also assume that $f_i(\mathbf{x}) = g(\mathbf{x} - \mu_i)$ $(i = 1, 2)$ for some location parameters μ_1 and μ_2 and a common elliptically symmetric density function g satisfying $g(k\mathbf{x}) \geq g(\mathbf{x})$ for every $\mathbf{x}$ and $0 < k < 1$. Then, under the conditions assumed in Theorem 5.1, the average misclassification probabilities for the regression depth based linear classifier converges to the optimal Bayes error as the training sample size tends to infinity provided that the prior probabilities of the two classes are equal. Further, in the equal prior case, if the Bayes classifier is unique and $U(\alpha)$ has a unique maximizer, the same holds for the half space depth based classifier, and in this case both of these two depth based classifiers themselves converge almost surely to that Bayes classifier. When the prior probabilities are unequal, the above convergence results for depth based linear classifiers remain true for normally distributed populations with a common dispersion matrix but different mean vectors.*

Corollary 5.3 : *Suppose that the population distributions f_1 and f_2 both belong to the class of elliptically symmetric multivariate normal or Pearson type-VII distributions, and they are of the same form except possibly for their location and scatter parameters. Then, the average misclassification rate of the quadratic classifier constructed using regression depth converges to the optimal Bayes error, and the quadratic classifier itself converges almost surely to the optimal Bayes classifier as the training sample size grows to infinity.*

Recall that the probability density function of a d-dimensional elliptically symmetric Pearson type-VII distribution is given by

$$f(\mathbf{x}) = (\pi\nu)^{-d/2}\Gamma(M)/\Gamma(M - d/2)\ |\Sigma|^{-1/2}\{1 + \nu^{-1}(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{-M},$$

where μ and Σ are the location and the scatter parameters, $\nu > 0$ and $M > d/2$ (see e.g., Fang, Kotz and Ng, 1989). When $M = (\nu + d)/2$ and ν is an integer, the corresponding distribution is known as the multivariate t-distribution with ν d.f. In the special case $\nu = 1$, we get the multivariate Cauchy distribution. Because of the heavy tails of such multivariate distributions, the traditional linear and quadratic classifiers would not perform satisfactorily in discriminating among such distributions. However, the above theorem and

the corollaries imply that the depth based linear and quadratic classifiers can achieve good misclassification rates for distributions with exponential tails like multivariate normal as well as for multivariate Cauchy and other distributions having heavy polynomial tails.

We conclude this section by pointing out an important fact related to the asymptotic convergence results stated in this section. All of these results have been stated for the case when the dimension r of the space of projection does not vary with the sample size N. On the other hand, in some non-parametric discriminant analysis methods e.g., those based on support vector machines (see e.g., Vapnik, 1995, 1998) or neural nets (see e.g., Ripley, 1994, 1996), the dimension of the space of projection usually grows with the sample size. For the depth based method also one may allow this kind of flexibility with respect to the choice of the discriminating surface. It will be clear from the proofs given in the Section 5.8 that if r grows with N in such a way that for all positive values of C, we have $\sum_{N\geq 1} N^{2r} e^{-CN} < \infty$, the convergence results in (i) and (ii) in Theorem 5.1 hold good. For instance, if r grows at the rate of N^t for any $0 < t < 1$, these convergence results remain valid.

5.4 Data analytic implementation

As we have already observed in Section 5.1, maximization of $U_{\mathbf{n}}(\alpha)$ w.r.t. α requires the computation of the half-space depth of the origin in the data cloud formed by the r-dimensional vectors of differences $\mathbf{z}_{1i} - \mathbf{z}_{2j}$ $(i = 1, 2, \ldots, n_1;\ j = 1, 2, \ldots, n_2)$. It is a finite maximization problem (see e.g., Chaudhuri and Sengupta, 1993), however, maximization by complete enumeration would lead to computational complexity of the order $O(n_\circ^{2r})$, where $n_\circ = max\{n_1, n_2\}$. An algorithm due to Rousseeuw and Ruts (1996) can reduce the computational complexity to order $O(n_\circ^{2(r-1)} log\ n_\circ)$. Similarly, maximization of $\Delta_{\mathbf{n}}(\alpha, \beta)$ w.r.t. α and β has computational complexity $O(n_\circ{}^r log\ n_\circ)$. Rousseeuw and Struyf (1998) provided some algorithms for computing location depth and regression depth. Some other optimization algorithms for regression depth are also available in Rousseeuw and Hubert (1999) and in Christmann *et. al.*, (2002).

5.4.1 Optimization of $U_{\mathbf{n}}(\alpha)$ and $\Delta_{\mathbf{n}}(\alpha, \beta)$

Recall from Section 5.2 that the maximization of $U_{\mathbf{n}}(\alpha)$ can be restricted to the α's with $\|\alpha\| = 1$ and the minimization of $\Delta(\alpha, \beta)$ can be restricted to (α, β)'s with $\|(\alpha, \beta)\| = 1$. However, since the order of the computational complexity increases rapidly with the dimension r, exact optimization of $U_{\mathbf{n}}(\alpha)$ and $\Delta_{\mathbf{n}}(\alpha, \beta)$ is not feasible for high dimensional problems, and there one can only resort to some approximate optimization. In this chapter, we have used a procedure, where the indicator functions appearing in the expressions for $U_{\mathbf{n}}$

and Δ_n get approximated by suitably chosen smooth functions. This approximation allows us to use the derivatives to find out the direction of steepest ascent/descent of the objective function to be optimized. A very simple approximation for the indicator function $I(x > 0)$ is the logistic function $1/(1 + e^{-tx})$ with large positive t. Clearly, a value of t, which is not large enough, will make the approximation very inaccurate. On the other hand, a very large value of t will make the approximation quite accurate but will make the numerical optimization using steepest ascent/descent numerically rather unstable. We have observed that a greater degree of numerical stability in the optimization algorithm can be achieved even for fairly large values of t if all measurement variables are standardized before the approximations are done. In all our numerical studies reported in the next two sections, we have found that if we use $5 \leq t \leq 10$ after standardizing the measurement variables, the average misclassification rates for the resulting procedures remain more or less same, and they are fairly low. Consequently, we have reported the best values obtained in that range. For linear discriminant analysis in the bivariate case, where exact computation of $U_n(\alpha)$ and $\Delta_n(\alpha, \beta)$ is easy, we have compared the performance of the exact and the approximate versions of these depth based classification methods and found them to achieve nearly similar average misclassification rates. In order to cope up with the problem of possible presence of several local minima, we have always run our approximate versions of the optimization algorithms a few times starting from different random initial points.

In the case of half-space depth based classifier, after estimating α, we need to estimate β from the training sample. This is done by enumerating the order statistics of the projected data points $\hat{\alpha}_H' z_{1i}$ and $\hat{\alpha}_H' z_{2j}$ $(1 \leq i \leq n_1, 1 \leq j \leq n_2)$ along the estimated direction $\hat{\alpha}_H$. Fortunately, since we use linear projections, the computational complexity in obtaining the estimate $\hat{\beta}_H$ does not increase with the dimension r.

5.4.2 Generalization of the procedure for multi-class problems

In a J-class $(J > 2)$ problem, to arrive at the final decision, one can use the method of majority voting (see e.g., Friedman, 1996), where binary classification is performed for each of the $\binom{J}{2}$ pairs of classes, and then an observation is assigned to the population which has the maximum number of votes. However, as we have mentioned earlier, this voting method may lead to some regions of uncertainty where more than one population can have the maximum number of votes. For instance, in a three class problem we may have a circular situation where each of the classes can have exactly one vote. When such situations occur, we can use the method of pairwise coupling as given in Hastie and Tibshirani (1998). Pairwise coupling is a method for combining the posterior weights of different populations obtained in different pairwise classifications. Recall that in our case, for any pairwise classification, an observation x is classified depending on the sign of $\alpha' z + \beta$. So, if σ is

some monotonically increasing function on the real line satisfying $0 \leq \sigma(x) \leq 1$, $\sigma(0) = 0.5$ and $\sigma(-x) = 1-\sigma(x)$ for every $x \in R$, we can use $\sigma(\boldsymbol{\alpha}'\mathbf{z}+\beta)$ as a measure of the strength in favor of the class determined by the inequality $\boldsymbol{\alpha}'\mathbf{z}+\beta > 0$. This can be taken as some kind of an estimate for the posterior weight in favor of that class in our pairwise comparison. Similarly, $1-\sigma(\boldsymbol{\alpha}'\mathbf{z}+\beta)$ can be used as an estimate for the posterior weight for the class determined by the inequality $\boldsymbol{\alpha}'\mathbf{z}+\beta < 0$. After one gets these posterior weights from pairwise comparisons, coupling can be conveniently used to get the combined weights for each of the J populations, and the observation can be classified to the population having highest combined posterior weight.

In all our numerical studies reported in the following sections of this chapter, we have taken σ to be the simple logistic function i.e., $\sigma(x) = 1/(1+e^{-x})$. We have applied pairwise coupling only for those rare observations, which did not get classified uniquely by the method of majority voting.

5.5 Results on simulated examples

In this section, we report our findings on some simulation studies that illustrate the performance of depth based classifiers as compared with traditional linear and quadratic classifiers. In all our simulated examples, we have restricted ourselves to two-class problems only, where the priors for both the populations are taken to be equal.

We first consider spherically symmetric multivariate normal and Cauchy distributions (with $\Sigma = \mathbf{I}$), which differ only in their location parameters. To make our examples simpler, we choose the location parameters $\boldsymbol{\mu}_1 = (0, 0, \ldots, 0)'$ and $\boldsymbol{\mu}_2 = (\mu, \mu, \ldots, \mu)$, where μ is taken to be only 1 and 2 in our experiments. For each of these examples, we generated 100 sets of training samples taking equal number of observations (either 50 or 100) from both the classes, and we used 2000 observations to form each test set. Average test set misclassification probabilities and their standard errors over these 100 simulation runs are reported in Tables 5.1 and 5.2. Optimal Bayes errors are also given to facilitate the comparison. For two-dimensional problems, we present the results for the depth based classifiers based on the exact and the approximate computation of the linear classifiers, and they don't seem to have significantly different performance. This is very encouraging as the approximate algorithms run very fast even for fairly high dimensional problems. From now on we will write H-depth to denote the half-space depth and R-depth to denote the regression depth in all the tables and subsequent discussions in this chapter.

		Bayes risk	n	*LDA*	H-depth		R-depth	
					Exact	Approx.	Exact	Approx.
Normal	$\mu = 1$	23.98	50	24.40 (0.10)	25.21 (0.14)	25.19 (0.15)	25.44 (0.15)	25.42 (0.13)
			100	24.21 (0.10)	24.80 (0.10)	24.72 (0.13)	25.11 (0.12)	24.88 (0.13)
	$\mu = 2$	7.87	50	8.23 (0.07)	8.96 (0.11)	8.91 (0.11)	9.15 (0.15)	8.99 (0.11)
			100	8.11 (0.07)	8.56 (0.11)	8.48 (0.11)	8.62 (0.09)	8.57 (0.09)
Cauchy	$\mu = 1$	30.40	50	43.81 (0.95)	32.45 (0.26)	32.51 (0.24)	32.45 (0.25)	32.50 (0.27)
			100	41.95 (0.98)	31.78 (0.15)	31.80 (0.15)	31.77 (0.15)	31.59 (0.14)
	$\mu = 2$	19.58	50	32.02 (1.34)	21.11 (0.19)	21.22 (0.19)	21.01 (0.16)	20.92 (0.15)
			100	33.19 (1.31)	20.83 (0.15)	20.77 (0.14)	20.60 (0.13)	20.43 (0.11)
Perturbed Normal	$\mu = 1$	22.71	50	50.75 (0.53)	29.15 (0.15)	28.96 (0.15)	29.21 (0.16)	29.20 (0.16)
			100	50.28 (0.53)	28.55 (0.12)	28.65 (0.13)	28.66 (0.13)	28.70 (0.12)
	$\mu = 2$	7.46	50	49.69 (0.25)	13.39 (0.10)	13.33 (0.11)	13.52 (0.11)	13.29 (0.09)
			100	50.41 (0.36)	12.98 (0.09)	12.97 (0.09)	13.02 (0.09)	12.87 (0.08)

Table 5.1 : Results on linear discrimination : average misclassification rates (in percentages) with standard errors (dimension 2).

			d=3		d=4	
			$\mu = 1$	$\mu = 2$	$\mu = 1$	$\mu = 2$
Normal	Bayes risk		19.32	4.16	15.87	2.28
	n=50	*LDA*	20.65 (0.16)	4.76 (0.07)	17.32 (0.15)	2.72 (0.06)
		H-depth	21.00 (0.15)	5.09 (0.10)	17.57 (0.15)	3.59 (0.10)
		R-depth	21.22 (0.16)	5.18 (0.10)	18.04 (0.18)	3.31 (0.08)
	n=100	*LDA*	19.64 (0.09)	4.28 (0.05)	16.33 (0.09)	2.42 (0.03)
		H-depth	20.05 (0.12)	4.75 (0.07)	16.78 (0.12)	3.06 (0.07)
		R-depth	20.37 (0.12)	4.73 (0.07)	17.14 (0.13)	2.90 (0.06)
Cauchy	Bayes risk		27.29	16.67	24.98	14.73
	n=50	*LDA*	40.15 (0.87)	26.96 (1.14)	37.36 (0.81)	23.85 (0.79)
		H-depth	30.03 (0.26)	18.79 (0.19)	28.50 (0.25)	17.43 (0.19)
		R-depth	29.68 (0.23)	18.38 (0.19)	27.59 (0.23)	16.87 (0.18)
	n=100	*LDA*	39.21 (0.90)	27.67 (0.98)	37.21 (0.87)	26.98 (1.19)
		H-depth	29.22 (0.18)	18.03 (0.14)	27.35 (0.22)	16.65 (0.13)
		R-depth	28.87 (0.15)	17.61 (0.12)	26.93 (0.16)	16.25 (0.11)
Perturbed Normal	Bayes risk		18.32	3.95	15.04	2.15
	n=50	*LDA*	50.28 (0.23)	50.14 (0.15)	49.99 (0.15)	50.00 (0.12)
		H-depth	24.60 (0.13)	10.08 (0.11)	21.87 (0.15)	8.52 (0.12)
		R-depth	24.89 (0.17)	9.99 (0.09)	22.28 (0.17)	8.46 (0.11)
	n=100	*LDA*	49.71 (0.27)	50.04 (0.15)	49.98 (0.13)	49.96 (0.11)
		H-depth	24.23 (0.11)	9.65 (0.08)	21.02 (0.11)	8.00 (0.06)
		R-depth	24.52 (0.12)	9.48 (0.06)	21.26 (0.12)	7.85 (0.07)

Table 5.2 : Results on linear discrimination : average misclassification rates (in percentages) with standard errors (dimensions 3 and 4).

Since the optimal Bayes rules are linear in the case of above mentioned spherically symmetric populations, good linear classifiers are expected to have error rates very close to the optimal Bayes risk. When the underlying distributions are multivariate normal, the traditional linear discriminant analysis (*LDA*) performed very well as one would expect.

However, the depth based methods also had a decent and comparable performance. But, in the case of multivariate Cauchy distribution, the depth based classifiers clearly outperformed LDA, and their performance is far closer to the optimal Bayes classifier than that of LDA.

Further, the performance of LDA was observed to fall drastically, when we added a small perturbation to the normally distributed data. We tried examples, where data in class-2 were taken to be normally distributed as before, and 10% of the observations in class-1 were replaced by observations having $N(10\mu_2, \mathbf{I})$ distributions. LDA in this case performed very poorly compared to both of the depth based classification techniques. Notice that the optimal Bayes rule is not linear in this case. Hence, none of the linear classifiers could achieve the accuracy of the optimal Bayes classifier.

			d=2		d=3		d=4	
			$\mu=1$	$\mu=2$	$\mu=1$	$\mu=2$	$\mu=1$	$\mu=2$
Normal	Bayes risk		22.03	13.31	16.62	8.34	12.89	5.37
	n=50	QDA	23.07 (0.10)	13.75(0.09)	17.97 (0.10)	9.13 (0.07)	14.80 (0.13)	6.41 (0.08)
		H-depth	25.08 (0.21)	14.99 (0.28)	20.40 (0.22)	11.18 (0.19)	17.13 (0.25)	8.65 (0.20)
		R-depth	25.09 (0.20)	15.35 (0.18)	20.31 (0.20)	10.99 (0.18)	16.99 (0.21)	8.30 (0.17)
	n=100	QDA	22.55 (0.10)	13.53 (0.07)	17.36 (0.09)	8.67 (0.06)	13.86 (0.09)	5.80 (0.06)
		H-depth	23.61 (0.14)	14.24 (0.11)	18.69 (0.15)	10.05 (0.13)	15.22(0.13)	7.32 (0.13)
		R-depth	23.85 (0.14)	14.58 (0.11)	18.73 (0.14)	9.94 (0.12)	15.18 (0.13)	7.17 (0.11)
Cauchy	Bayes risk		30.92	22.97	28.36	19.84	26.49	17.76
	n=50	QDA	46.63 (0.39)	45.86 (0.55)	46.13 (0.43)	43.59 (0.58)	45.08 (0.44)	43.47 (0.65)
		H-depth	34.70 (0.24)	26.12 (0.19)	32.58 (0.22)	23.43 (0.21)	31.17 (0.23)	21.36 (0.24)
		R-depth	34.29 (0.26)	26.11 (0.20)	33.48 (0.27)	23.45 (0.20)	31.05 (0.23)	22.12 (0.25)
	n=100	QDA	48.08 (0.32)	46.90 (0.34)	47.50 (0.32)	46.89 (0.39)	46.29 (0.30)	44.84 (0.41)
		H-depth	33.24 (0.16)	25.02 (0.14)	31.10 (0.18)	22.22 (0.16)	29.36 (0.19)	20.49 (0.14)
		R-depth	33.30 (0.19)	24.96 (0.17)	31.35 (0.19)	22.47 (0.17)	29.52 (0.20)	20.55 (0.14)
Perturbed Normal	Bayes risk		21.36	12.90	16.10	8.06	12.46	5.20
	n=50	QDA	38.42 (0.49)	28.62 (0.57)	28.95 (0.31)	17.80 (0.35)	23.61 (0.23)	13.50 (0.26)
		H-depth	25.85 (0.24)	15.01 (0.16)	22.75 (0.30)	12.71 (0.26)	20.88 (0.28)	11.77 (0.24)
		R-depth	28.23 (0.28)	16.81 (0.26)	24.70 (0.24)	14.58 (0.19)	21.26 (0.20)	12.34 (0.20)
	n=100	QDA	39.08 (0.33)	29.71 (0.42)	28.32 (0.19)	17.70 (0.22)	22.78 (0.16)	12.73 (0.21)
		H-depth	24.85 (0.19)	14.43 (0.15)	20.76 (0.23)	10.69 (0.19)	18.73 (0.21)	9.49 (0.23)
		R-depth	26.88 (0.22)	15.66 (0.23)	22.61 (0.18)	12.33 (0.23)	19.82 (0.18)	11.02(0.16)

Table 5.3 : Results on quadratic discrimination : average misclassification rates (in percentages) with standard errors.

Results obtained in the case of quadratic discrimination are reported in Table 5.3, and here too we find similar behavior of the competing classifiers as in the case of linear discriminant analysis. We used the same mean vectors as before but took two different scatter matrices for the two competing populations (with distributions normal or Cauchy), namely $\Sigma_1 = \mathbf{I}$ and $\Sigma_2 = 4\mathbf{I}$. The traditional quadratic discriminant analysis (QDA) performed well in discriminating multivariate normal populations but its performance turned out to be very poor in the case of multivariate Cauchy populations as well as multivariate perturbed normal populations. The two depth based quadratic classifiers, on the other hand, showed decent performance in the case of normally distributed data, and had average misclassification rates much closer to the optimal Bayes risks than the error rates of QDA in the case of multivariate Cauchy and perturbed normal distributions.

In all these simulated examples, the performance of the two depth based classifiers

were fairly similar except for quadratic classification in the case of perturbed normal distribution, where the H-depth based classifier had a small edge over the R-depth based classifier for all sample sizes and all dimensions.

5.6 Results from the analysis of benchmark data sets

We now investigate the performance of the depth based classifiers on six well known data sets. Apart from crab data, all these data sets have been used in the earlier chapters of this thesis. In the case of the first two data sets (the vowel recognition data-1 and the synthetic data), there are well-defined training and test sets. For them, we have reported the performance of different competing classifiers on those test sets. In each of the remaining four cases, we have divided the data randomly into two parts to form the training and the test samples. This random division is carried out 1000 times to generate 1000 different partitions for each data set. Average test set misclassification errors over these 1000 random partitions and their corresponding standard errors are reported in Table 5.4. In all the examples, sample proportions for different classes have been used as their prior probabilities.

Vowel data-1 : We begin with the vowel recognition data-1 described in Section 3.4. It contains bivariate observations from 10 different populations. On this data set, traditional LDA led to a test set error rate of 25.26 % (with a standard error of 2.38%), but using depth based linear classifiers, we were able to get significantly better results. The linear classifiers based on H-depth and R-depth could reduce the average misclassification probability to 20.72% (with a standard error of 2.22%) and 19.83% (with a standard error of 2.18%) respectively. Interestingly, as reported in Table 5.4, in the case of quadratic classifiers, the performance of the two depth based classification rules and that of the traditional QDA applied to the test set turned out to be fairly similar for this data.

Synthetic data : This is a data set consisting of 250 bivariate observations for the training set and 1000 observations for the test set coming from two competing populations, and it has already been discussed in the previous two chapters. The average misclassification rates of different classification methods applied on this test set are reported in Table 5.4. For linear as well as for quadratic classification, the error rates of the traditional and the depth based methods were fairly similar. Figures 5.2 and 5.3 show the performance of these linear and quadratic classifiers on the training and the test sets. For both of the linear and the quadratic classification, the estimated class boundaries for the traditional and the depth based classifiers were found to be almost identical.

Diabetes data : As we have mentioned in Chapters 3 and 4, this data set has 145 observations, which are distributed as 33, 36 and 76 in those three classes. Since it does not

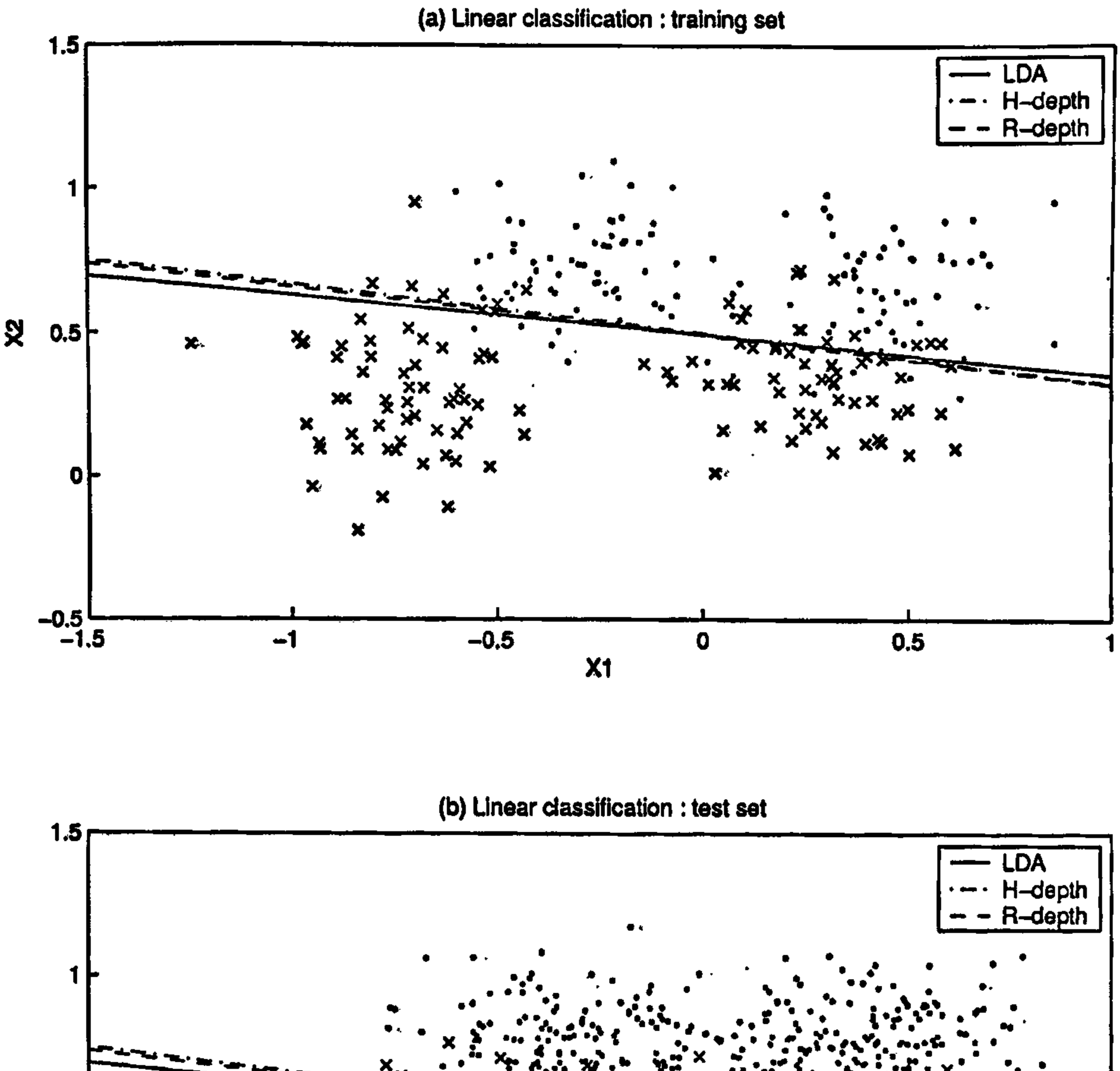
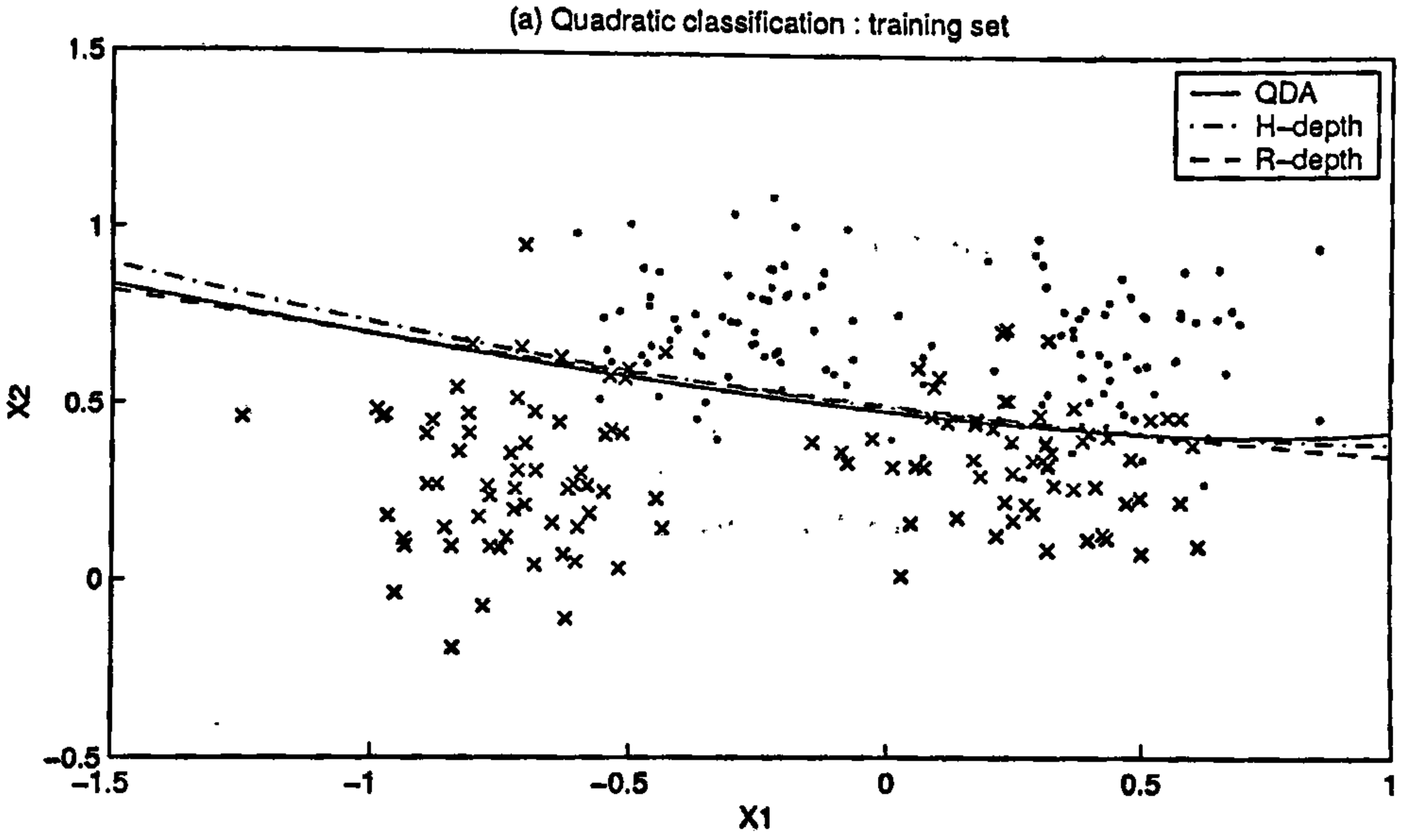


Figure 5.2 : Linear classification on synthetic data



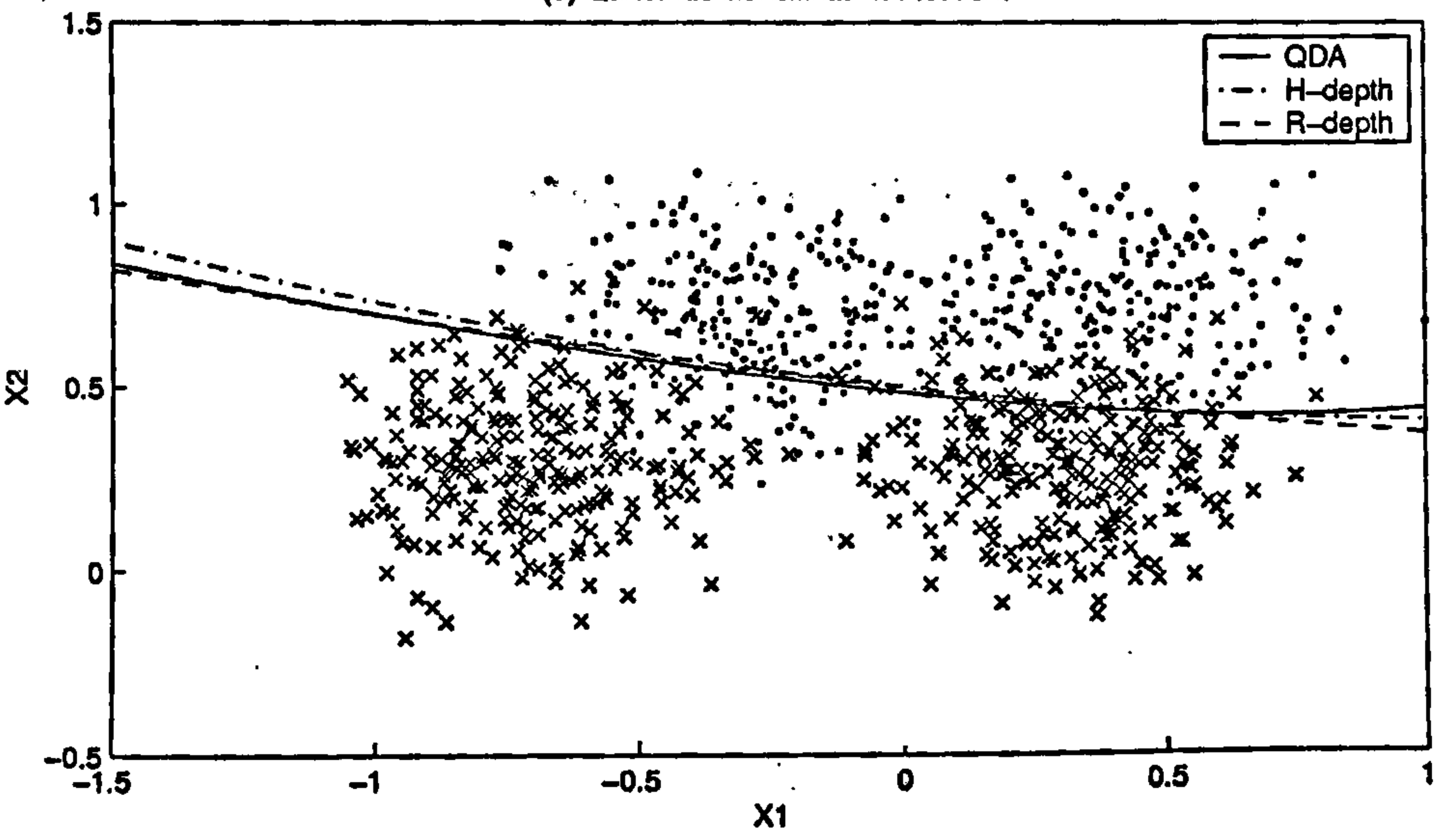


Figure 5.3 : Quadratic classification on synthetic data

have separate training and test sets, to evaluate the performance of different classifiers, we have partitioned it randomly to form the training and the test samples. Training samples are formed by taking 25 observations from each of the first two populations and 50 observations from the third, while the rest of the observations were used to form the corresponding test samples.

In this data set, the depth based procedures clearly outperformed traditional LDA and QDA. While traditional LDA showed an average misclassification rate of 11.12% (standard error = 0.07%), the H-depth and the R-depth based linear classifiers could reduce this error rate to 5.49% (standard error = 0.06%) and 6.12% (standard error = 0.06%), respectively. Though traditional QDA (error rate = 9.32%, standard error = 0.06%) performed better than traditional LDA, it is quite apparent from the figures in Table 5.4 that depth based quadratic classifiers had a clear edge over the traditional one.

Bio-medical data : Like diabetes data, this data set has also been discussed in Chapter 4, and it does not have separate training and test samples. 100 observations from the first group and 50 from the second were chosen randomly to form each training sample, while the remaining observations were used to form the corresponding test samples. Here also, the depth based linear classifiers outperformed traditional LDA. As shown in Table 5.4, LDA had an error rate of 15.96% (with a standard error of 0.07%), while the H-depth and the R-depth based classifiers could reduce it up to 10.87% (standard error = 0.07%) and 11.03% (standard error = 0.07%), respectively. Figures reported in Table 5.4 also indicate that even in the case of quadratic discriminant analysis, these depth based classifiers led to a significantly better performance than traditional QDA.

Crab data : Campbell and Mahon (1974) used this data set for morphological study on the rock crabs of the genus Leptograpsus. One species had been split into two new species, which were previously marked by colors 'orange' and 'blue'. As the preserved specimens lost their colors, it was hoped that the morphological study would help to classify the museum materials. This data set contains the information on 50 specimens of each sex of each of the species. For each specimen there are measurements on five different variables (body depth and four other measurements on carapace). We have randomly taken 40 observations from each of the four classes to form a training set while remaining observations have been used as the corresponding test sample. For this data set, the results reported in Table 5.4 show that the depth based classifiers and traditional LDA and QDA have comparable performance with depth based methods having a small edge over the traditional techniques.

Iris data : As the last example of this section, we consider the Iris data introduced in Section 2.5 of Chapter 2 and also discussed subsequently in Chapter 4. Here, we have randomly chosen 40 observations from each class to construct a training sample, and the remaining 30 observations have been used to form the corresponding test set. It is quite well

known that traditional LDA and QDA perform very well for this data set, and depth based classifiers are not expected to beat them in this case. However, the error rates reported in Table 5.4 show that both the linear and the quadratic versions of the depth based methods could produce a decent and comparable performance for this data set.

	Linear classification			Quadratic classification		
	LDA	H-depth	R-depth	QDA	H-depth	R-depth
Vowel data	25.26 (2.38)	20.72 (2.22)	19.83 (2.18)	19.83 (2.18)	19.22 (2.16)	19.53 (2.17)
Synthetic Data	10.80 (0.98)	10.70 (0.98)	10.30 (0.96)	10.20 (0.96)	10.70 (0.98)	11.00 (0.99)
Diabetes Data	11.12 (0.07)	5.49 (0.06)	6.12 (0.06)	9.32 (0.06)	6.57 (0.06)	7.09 (0.06)
Bio-medical Data	15.96 (0.07)	10.87 (0.07)	11.03 (0.07)	12.68 (0.06)	11.61 (0.07)	11.76 (0.06)
Crab Data	5.20 (0.06)	4.85 (0.06)	4.47 (0.06)	5.89 (0.06)	4.37 (0.06)	4.26 (0.06)
Iris Data	2.18 (0.07)	3.92 (0.10)	3.56 (0.10)	2.75 (0.09)	3.99 (0.11)	3.43 (0.10)

Table 5.4 : Results on benchmark data sets : average misclassification rates (in percentages) with standard errors.

5.7 Remarks and discussions

Use of data depth in discriminant analysis was first proposed by Liu (1990), where she suggested to classify an observation using its relative center-outward ranks with respect to different populations obtained using some depth function. Jornsten, Vardi and Zhang (2002) and Jornsten (2004) used that idea to develop nonparametric methods for clustering and classification based on L_1 depth (also known as spatial depth) function (see e.g., Vardi and Zhang, 2000; Serfling, 2002). However, for classifying a new observation, this classifier needs to calculate its depth with respect to different competing populations, and for that the full training sample have to be stored. On the other hand, H-depth and R-depth classifiers require less storage and computing time for classifying future observations, and at the same time they provide good lower dimensional view of class separability.

Both of traditional LDA and QDA are motivated by the assumption of normality of the data, and as we have amply demonstrated in preceding sections, violations in this assumption may lead to rather poor performance of these traditional methods. More recent methods like regularized discriminant analysis due to Friedman (1989) and logistic discriminant analysis (see e.g., Hand, 1981; Hastie *et. al.*, 2001) are also motivated by specific distributional models for the data. The depth based classifiers, on the other hand, are totally distribution free in nature, and they use only the empirical geometry of the data cloud to estimate the optimal separating surface for the competing classes. Traditional LDA, QDA as well as regularized discriminant analysis use the first and the second order moments of the training sample to construct the discrimination rule. This makes these methods highly sensitive to outliers and extreme values. On the other hand, use of half-space and regression depths in the construction of the classifiers makes the discriminant

functions more robust against the presence of possible outliers in the case of heavy-tailed distributions.

For nonlinear classification, the depth based methods project the observations into a space of functions to find a separating hyperplane there. Well-known nonparametric methods like those based on neural nets (see e.g., Ripley, 1994, 1996) and support vector machines (SVM) (see e.g., Vapnik, 1995, 1998, Scholkopf, Burges and Smola, 1999; Cristianini and Shawe-Taylor, 2000) also adopt a similar strategy for nonlinear classification. However, instead of minimizing the empirical misclassification rates as it is done in the case of depth based methods, these classifiers are formed by minimizing some smooth penalty functions. Other techniques like flexible discriminant analysis due to Hastie, Tibshirani and Buja (1994) and the classifier recently proposed by Zhu and Hastie (2003) also optimize some smooth cost or likelihood type functions to determine the discriminant function.

We conclude this section with an illustrative example taken from Christmann (2002). This is a simulated example on a four class problem, where the classes are completely separated (see figure 5.4). An observation (x_1, x_2) in the square $[-1, 1] \times [-1, 1]$ is assigned to class-1 if $x_2 - x_1 > 0.75$ and to class-2 if $x_1^2 + x_2^2 \leq 0.15$. An observation (x_1, x_2) satisfying $x_2 - x_1 \leq 0.75$ and $x_1^2 + x_2^2 > 0.15$ is assigned to class-3 or class-4 depending on whether $x_1^2 + x_2^2 \leq 0.60$ or > 0.60, respectively.

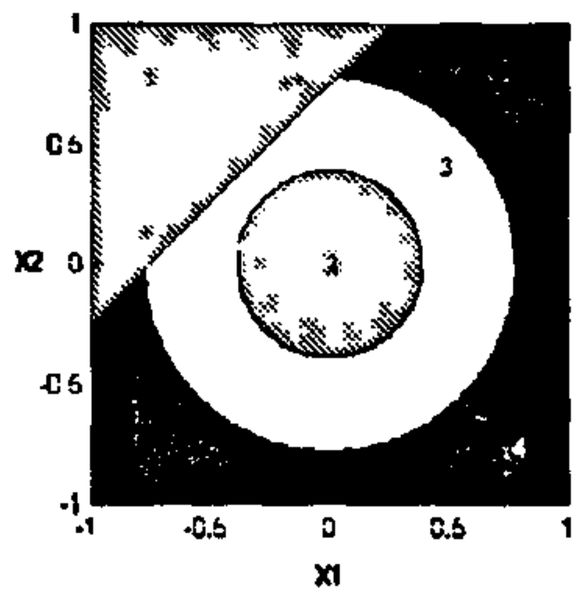


Figure 5.4 : A four class problem

Christmann (2002) generated 250 different training samples each of size 700 and test samples each of size 300 to compare the performance of SVM with that of traditional QDA. In this example, SVM (with radial basis function) produced a much higher average error rate of 36% than QDA having average misclassification rate of 20.9%. We have generated 250 samples of the same sizes as used by Christmann (2002) to compare the performance of the depth based classifiers. In our experiment, QDA produced almost similar performance (error rate = 20.72 %) as reported by Christmann (2002), but the quadratic versions of both of the depth based classifiers performed quite well. H-depth and R-depth based classifiers on this example led to an average test set error rate of 1.58% (standard error = 0.03%) and 2.81% (standard error = 0.17%), respectively.

5.8 Proofs and mathematical details

In order to prove Theorem 5.1, we will need the following result, which follows directly from the proof of Lemma A of Serfling (1980), p. 200.

Result 5.1 : If Y is a bounded random variable with $E(Y) = \mu$ and $P(0 \leq Y \leq 1) = 1$, then

$$E\{e^{s(Y-\mu)}\} \leq e^{s^2/8} \text{ for any } s > 0.$$

Proof of Theorem 5.1 : (i) $U_{\mathbf{n}}(\alpha)$ is a generalized U-statistic (see e.g., Serfling, 1980) having the bounded kernel function $q(\alpha' \mathbf{z}_1, \alpha' \mathbf{z}_2) = I\{\alpha' \mathbf{z}_1 > \alpha' \mathbf{z}_2\}$ $(0 \leq q \leq 1)$. Now, without loss of generality, let us assume that $n_1 \leq n_2$ and define

$$W(i_1, i_2, \ldots, i_{n_1}) = {n_1}^{-1} \sum_{j=1}^{n_1} q(\alpha' \mathbf{z}_{1j}, \alpha' \mathbf{z}_{2i_j})$$

for some permutation $(i_1, i_2, \ldots, i_{n_1})$ of n_1 objects from $\{1, 2, \ldots, n_2\}$. For this definition of W, $U_{\mathbf{n}}(\alpha)$ can be expressed as

$$U_{\mathbf{n}}(\alpha) = \frac{(n_2 - n_1)!}{n_2!} \sum_{(i_1, i_2, \ldots, i_{n_1}) \in \mathcal{P}} W(i_1, i_2, \ldots, i_{n_1}),$$

where $\mathcal{P}$ denotes the set of all possible permutations $(i_1, i_2, \ldots, i_{n_1})$ of the elements of the set $\{1, 2, \ldots, n_2\}$.

Now, using Jensen's inequality on the convex function e^x, we get

$$e^{sU_{\mathbf{n}}(\alpha)} \leq \frac{(n_2 - n_1)!}{n_2!} \sum_{(i_1, i_2, \ldots, i_{n_1}) \in \mathcal{P}} e^{sW(i_1, i_2, \ldots, i_{n_1})} \text{ for every } s > 0$$

$$\Rightarrow E\{e^{sU_{\mathbf{n}}(\alpha)}\} \leq E\{e^{sW(i_1, i_2, \ldots, i_{n_1})}\} = \left[E\{e^{sq(\alpha' \mathbf{z}_{11}, \alpha' \mathbf{z}_{21})/n_1}\}\right]^{n_1}$$

(using the fact that the terms in the sum defining W are independent and identically distributed)

$$\Rightarrow E\left\{e^{s[U_{\mathbf{n}}(\alpha) - U(\alpha)]}\right\} \leq \left[E\{e^{s[q(\alpha' \mathbf{z}_{11}, \alpha' \mathbf{z}_{21}) - U(\alpha)]/n_1}\}\right]^{n_1} = \{\Psi_q(s/n_1)\}^{n_1}, \text{ say.}$$

Now, it is quite easy to see that

$$E\{U_{\mathbf{n}}(\alpha)\} = E\{W(i_1, i_2, \ldots, i_{n_1})\} = E\{q(\alpha' \mathbf{z}_1, \alpha' \mathbf{z}_2)\} = P\{\alpha' \mathbf{z}_{11} > \alpha' \mathbf{z}_{21}\} = U(\alpha), \text{ and}$$

using Result 5.1, we get for any $t > 0$,

$$P\{U_{\mathbf{n}}(\alpha) - U(\alpha) \geq t\} \leq E\{e^{s[U_{\mathbf{n}}(\alpha) - U(\alpha) - t]}\} \leq e^{-st}\{\Psi_q(s/n_1)\}^{n_1} \leq e^{-st + \frac{s^2}{8n_1}}.$$

Minimizing the above expression with respect to s, we get $P\{U_{\mathbf{n}}(\alpha) - U(\alpha) \geq t\} \leq e^{-2n_1t^2}$. Using similar arguments, it can be shown that for any positive t, $P\{U_{\mathbf{n}}(\alpha) - U(\alpha) \leq -t\} \leq e^{-2n_1t^2}$. Combining these two results, we obtain

$$P\left\{\left|U_{\mathbf{n}}(\alpha) - U(\alpha)\right| \geq t\right\} \leq 2e^{-2n_1t^2} \quad \text{for every } t > 0.$$

The above result can also be obtained as a special case of Theorem 2.1 of Serfling (1992). Now, the set of hyperplanes in $H = \{\mathbf{y} : \alpha'\mathbf{y} = 0\}$ in R^r, which pass through the origin has VC dimension r (see e.g., Pollard, 1984; Vapnik, 1998). So, the sets of the form $\{\mathbf{y} : \alpha'\mathbf{y} > 0\}$ has a polynomial discrimination with r being the degree of the polynomial. Therefore, using the results on probability inequalities on such sets (see e.g., Vapnik and Chervonenkis, 1971; Pollard, 1984; Vapnik 1998), we get

$$P\left\{\sup_{\alpha}|U_{\mathbf{n}}(\alpha) - U(\alpha)| > t\right\} < 2\,(n_1n_2)^r e^{-2n_1t^2} \quad \text{for every } t > 0.$$

Now, using the fact that $n_1/N \to \lambda$ $(0 < \lambda < 1)$ as $N \to \infty$, and $\sum_{N\geq 1} N^{2r}e^{-CN} < \infty$ for any $C > 0$, it follows from Borel Cantelli Lemma that $\sup_{\alpha}|U_{\mathbf{n}}(\alpha) - U(\alpha)| \to 0$ almost surely as $N \to \infty$.

Let $\hat{\alpha}_H$ be a maximizer of $U_{\mathbf{n}}(\alpha)$ and α_H^* be that of $U(\alpha)$ (not necessarily unique). Now, we have

$$|U_{\mathbf{n}}(\hat{\alpha}_H) - U(\hat{\alpha}_H)| \xrightarrow{a.s} 0 \quad \text{and } |U_{\mathbf{n}}(\alpha_H^*) - U(\alpha_H^*)| \xrightarrow{a.s} 0 \quad \text{as } N \to \infty.$$

Again, from the definition of $\hat{\alpha}_H$ and α_H^*, $U(\alpha_H^*) \geq U(\hat{\alpha}_H)$ and $U_{\mathbf{n}}(\hat{\alpha}_H) \geq U_{\mathbf{n}}(\alpha_H^*)$ for every $\mathbf{n}$. Hence, $|U_{\mathbf{n}}(\hat{\alpha}_H) - \max_{\alpha} U(\alpha)| = |U_{\mathbf{n}}(\hat{\alpha}_H) - U(\alpha_H^*)| \xrightarrow{a.s} 0$ as $N \to \infty$. Consequently, $|U(\hat{\alpha}_H) - \max_{\alpha} U(\alpha)| \xrightarrow{a.s} 0$ as $N \to \infty$.

(ii) For some fixed α and β, $\frac{1}{n_1}\sum_{i=1}^{n_1} I\{\alpha'\mathbf{z}_{1i} + \beta < 0\}$ is an average of i.i.d. bounded random variables. Therefore, from Hoeffding's (see Hoeffding, 1963) inequality, we have

$$P\left\{\left|\frac{1}{n_1}\sum_{i=1}^{n_1} I\{\alpha'\mathbf{z}_{1i} + \beta < 0\} - P\{\alpha'\mathbf{z}_{11} + \beta < 0\}\right| > \epsilon/2\right\} < 2e^{-n_1\epsilon^2/2} \quad \text{for every } \epsilon > 0.$$

Similarly, $P\left\{\left|\frac{1}{n_2}\sum_{i=1}^{n_2} I\{\alpha'\mathbf{z}_{2i} + \beta > 0\} - P\{\alpha'\mathbf{z}_{21} + \beta > 0\}\right| > \epsilon/2\right\} < 2e^{-n_2\epsilon^2/2}$ for every $\epsilon > 0$.

$$\begin{aligned}
\Rightarrow P\{|\Delta_{\mathbf{n}}(\alpha,\beta) - \Delta(\alpha,\beta)| > \epsilon\} &< P\left\{\left|\tfrac{1}{n_1}\sum_{i=1}^{n_1} I\{\alpha'\mathbf{z}_{1i} + \beta < 0\} - P\{\alpha'\mathbf{z}_{11} + \beta < 0\}\right| > \epsilon/2\right\} \\
&+ P\left\{\left|\tfrac{1}{n_2}\sum_{i=1}^{n_2} I\{\alpha'\mathbf{z}_{2i} + \beta > 0\} - P\{\alpha'\mathbf{z}_{21} + \beta > 0\}\right| > \epsilon/2\right\} \\
&< 2(e^{-n_1\epsilon^2/2} + e^{-n_2\epsilon^2/2}).
\end{aligned}$$

Now, using similar arguments on VC dimension of hyperplanes in R^r as before and using the results (see e.g., Pollard, 1984) on sets having polynomial discrimination, we get

$$P\left\{\sup_{\alpha,\beta}|\Delta_{\mathbf{n}}(\alpha,\beta)-\Delta(\alpha,\beta)|>\epsilon\right\}<2(n_1+n_2)^{r+1}(e^{-n_1\epsilon^2/2}+e^{-n_2\epsilon^2/2}).$$

Then, using the fact that $\sum_{N\geq 1} N^{r+1}e^{-CN}<\infty$ for any $C>0$, it follows from Borel Cantelli Lemma that $\sup_{\alpha,\beta}|\Delta_{\mathbf{n}}(\alpha,\beta)-\Delta(\alpha,\beta)|\to 0$ almost surely as $N\to\infty$.

Following similar arguments as used in the end of the proof of (i), it is now easy to verify that $|\Delta(\hat{\alpha}_R,\hat{\beta}_R)-\min_{\alpha,\beta}\Delta(\alpha,\beta)|\to 0$ and $|\Delta_{\mathbf{n}}(\hat{\alpha}_R,\hat{\beta}_R)-\min_{\alpha,\beta}\Delta(\alpha,\beta)|\to 0$ almost surely as $N\to\infty$.

Let us next assume that the maximizer α_H^* of $U(\alpha)$ is unique. We have already shown $U(\hat{\alpha}_H)$ converges to $U(\alpha_H^*)$ as $N\to\infty$ on a set of probability one. Consequently, on the same set, if $\hat{\alpha}_H$ converges, it has to converge to α_H^* in view of the uniqueness of α_H^* and the continuity of the function $U(\alpha)$. Since $\hat{\alpha}_H$ always lies in the compact surface of the unit ball in R^m (see Sections 5.2.1 and 5.4.1), any subsequence of the sequence of this estimate will have a further convergent subsequence converging to α_H^* on that set of probability one. Hence, $\hat{\alpha}_H$ must converge to α_H^* almost surely.

Next, let (α_R^*,β_R^*) be the unique minimizer of $\Delta(\alpha,\beta)$. Since we have already shown $\Delta(\hat{\alpha}_R,\hat{\beta}_R)$ converges to $\Delta(\alpha_R^*,\beta_R^*)$ almost surely, using arguments which are virtually same as those above, it follows that as $N\to\infty$, $(\hat{\alpha}_R,\hat{\beta}_R)\xrightarrow{a.s}(\alpha_R^*,\beta_R^*)$.

Proof of Corollary 5.1 : In Theorem 5.1, we have proved that $|\Delta(\hat{\alpha}_R,\hat{\beta}_R)-\min_{\alpha,\beta}\Delta(\alpha,\beta)|\to 0$ almost surely as $N\to\infty$. Note that $\Delta(\hat{\alpha}_R,\hat{\beta}_R)$ is the conditional average misclassification probability for a future observation given the current training sample. Taking expectation of $\Delta(\hat{\alpha}_R,\hat{\beta}_R)$ over the current training sample, the proof of this corollary follows by a simple application of Dominated Convergence Theorem using the fact that Δ is a function bounded between 0 and 1.

Lemma 5.1 : Suppose that the population densities f_1 and f_2 of the two competing classes are elliptically symmetric with a common scatter matrix Σ. Also assume that $f_i(\mathbf{x})=g(\mathbf{x}-\mu_i)$ $(i=1,2)$ for some location parameters μ_i and a common elliptically symmetric density function g satisfying $g(k\mathbf{x})\geq g(\mathbf{x})$ for every $\mathbf{x}$ and $0<k<1$. Further, assume that the prior probabilities of the two competing classes are equal. Then,

(i) there exists an optimal Bayes classifier, which is linear and

(ii) $\alpha=\Sigma^{-1}(\mu_1-\mu_2)$ is a maximizer of $U(\alpha)$ as well as a minimizer $\Delta(\alpha,\beta)$ for a proper choice of β.

Proof of Lemma 5.1 : (i) Because of elliptic symmetry with location shift, the density functions f_1 and f_2 can be expressed as

$$f_1(\mathbf{x}) = C_d|\Sigma|^{-1/2}\eta\{(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1)\} \text{ and } f_2(\mathbf{x}) = C_d|\Sigma|^{-1/2}\eta\{(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)\},$$

where C_d is a constant that depends on dimension d, and η is a monotonically decreasing function on $[0,\infty)$. Now, in the equal prior case, an optimum Bayes rule classifies an observation to class-1 if and only if

$$\begin{aligned} f_1(\mathbf{x}) \geq f_2(\mathbf{x}) &\Leftrightarrow (\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1) \leq (\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2) \\ &\Leftrightarrow (\mu_1-\mu_2)'\Sigma^{-1}\mathbf{x} \geq \tfrac{1}{2}\left[\mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2\right]. \end{aligned}$$

This proves that an optimal linear classifier is a Bayes classifier and $\alpha = \Sigma^{-1}(\mu_1-\mu_2)$ is a minimizer of $\Delta(\alpha,\beta)$ with a proper choice of β.

(ii) As the distributions have a common elliptically symmetric form with location parameters μ_1 and μ_2 and common scatter matrix Σ, their characteristic functions are of the form

$$\Psi_{f_1}(\mathbf{t}) = e^{i\mathbf{t}'\mu_1}\lambda(\mathbf{t}'\Sigma\mathbf{t}) \text{ and } \Psi_{f_2}(\mathbf{t}) = e^{i\mathbf{t}'\mu_2}\lambda(\mathbf{t}'\Sigma\mathbf{t})$$

for some common scalar function λ. Now define $Y = \alpha'\{(\mathbf{X}_1-\mathbf{X}_2)-(\mu_1-\mu_2)\}/(\alpha'\Sigma\alpha)^{1/2}$, where $\mathbf{X}_1 \sim f_1$ and $\mathbf{X}_2 \sim f_2$. It is easy to see that the characteristic function of Y is given by $\Psi_Y(t) = \{\lambda(t^2)\}^2$. Clearly, the distribution of Y is symmetric about 0, and it is free from population parameters like the μ's and the Σ. Therefore, $P\{\alpha'(\mathbf{X}_1-\mathbf{X}_2) > 0\}$ can be expressed as

$$P\{\alpha'(\mathbf{X}_1-\mathbf{X}_2) > 0\} = F_Y\left([\{\alpha'(\mu_1-\mu_2)\}^2/\alpha'\Sigma\alpha]^{1/2}\right),$$

where F_Y is the c.d.f. of the distribution of Y. So, $P\{\alpha'(\mathbf{X}_1-\mathbf{X}_2) > 0\}$ gets maximized for some α if that α maximizes $\{\alpha'(\mu_1-\mu_2)\}^2/\alpha'\Sigma\alpha$. This implies that $\alpha = \Sigma^{-1}(\mu_1-\mu_2)$ is a maximizer of $U(\alpha)$.

Proof of Corollary 5.2 : Lemma 5.1 implies that, under the given conditions, the linear classifier with $\alpha = \Sigma^{-1}(\mu_1-\mu_2)$ and $\beta = (\mu_2'\Sigma^{-1}\mu_2 - \mu_1'\Sigma^{-1}\mu_1)/2$ is a Bayes classifier. Consequently, it follows from Corollary 5.1 that the average misclassification error of the regression depth based linear classifier converges to the optimal Bayes risk. Further, when this Bayes classifier is unique, it follows from the second half of Theorem 5.1 that the regression depth based linear classifier itself converges almost surely to that Bayes classifier.

When $U(\alpha)$ has a unique maximizer $\alpha_H^* = \Sigma^{-1}(\mu_1-\mu_2)$ (e.g., when the distribution function F_Y in the proof of Lemma 5.1 is strictly increasing), it follows from Theorem 5.1 that $\hat{\alpha}_H$ converges almost surely to α^* as $N \to \infty$.

Let us now consider two independent random vectors $\mathbf{X}_1 \sim f_1$ and $\mathbf{X}_2 \sim f_2$ both of which being completely independent of the current training sample (i.e. they are like future observations). Using these random vectors define $Y_{1,\mathbf{n}} = \hat{\boldsymbol{\alpha}}_H' \mathbf{X}_1$, $Y_{2,\mathbf{n}} = \hat{\boldsymbol{\alpha}}_H' \mathbf{X}_2$, $Y_1 = \boldsymbol{\alpha}^{*\prime} \mathbf{X}_1$ and $Y_2 = \boldsymbol{\alpha}^{*\prime} \mathbf{X}_2$. Then, in view of almost sure convergence of $\hat{\boldsymbol{\alpha}}_H$ to $\boldsymbol{\alpha}^*$, we get $(Y_{1,\mathbf{n}}, Y_{2,\mathbf{n}}) \xrightarrow{L} (Y_1, Y_2)$ almost surely as $N \to \infty$. Since both of Y_1 and Y_2 are continuously distributed, and weak convergence to a continuous distribution implies uniform convergence, we have $\sup_{\boldsymbol{\beta}} |\Delta(\hat{\boldsymbol{\alpha}}_H, \boldsymbol{\beta}) - \Delta(\boldsymbol{\alpha}^*, \boldsymbol{\beta})| \to 0$ almost surely as $N \to \infty$.

On the other hand, from the proof of (ii) in Theorem 5.1, it is quite clear that $\sup_{\boldsymbol{\beta}} |\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H, \boldsymbol{\beta}) - \Delta(\hat{\boldsymbol{\alpha}}_H, \boldsymbol{\beta})| \to 0$ almost surely as $N \to \infty$. Hence, $\sup_{\boldsymbol{\beta}} |\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H, \boldsymbol{\beta}) - \Delta(\boldsymbol{\alpha}^*, \boldsymbol{\beta})| \to 0$ almost surely as $N \to \infty$.

It now follows from arguments similar to those used in the proof of Theorem 5.1, $|\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H) - \min_{\boldsymbol{\beta}} \Delta(\boldsymbol{\alpha}^*, \boldsymbol{\beta})| = |\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H) - \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \Delta(\boldsymbol{\alpha}, \boldsymbol{\beta})| \to 0$ almost surely as $N \to \infty$. Also, we must have $|\Delta_{\mathbf{n}}(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H) - \Delta(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H)| \to 0$ almost surely as $N \to \infty$. Hence, $\Delta(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H)$ converges almost surely to $\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \Delta(\boldsymbol{\alpha}, \boldsymbol{\beta})$, which is the Bayes risk in this case.

Once again, note that $\Delta(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H)$ is the conditional average misclassification probability for a future observation given the current training sample. Taking expectation of $\Delta(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H)$ over the current training sample we get the unconditional average misclassification probability of the linear classifier based on half-space depth. The proof of the convergence is now complete by a simple application of dominated convergence theorem using the fact that Δ is a function bounded between 0 and 1.

Now, to prove the almost sure convergence of the linear classifier based on half-space depth, we only need to show that $\hat{\boldsymbol{\beta}}_H$ converges almost surely to an appropriate constant. In order to prove that let us first recall a simple fact about the optimal Bayes classifier. In the equal prior case with two competing populations, it is easy to verify that the optimal Bayes risk is strictly smaller than 0.5 unless the two populations are statistically indistinguishable in the sense that they have identical distributions. We have already shown that $\Delta(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H)$ converges to the Bayes risk and $\hat{\boldsymbol{\alpha}}_H$ converges to $\boldsymbol{\alpha}^*$ as $N \to \infty$ on a set with probability one. So, on this set $\hat{\boldsymbol{\beta}}_H$ must remain bounded as otherwise in view of the convergence of $\hat{\boldsymbol{\alpha}}_H$ to $\boldsymbol{\alpha}^*$, $\Delta(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H)$ will converge to 0.5 along a subsequence along which $|\hat{\boldsymbol{\beta}}_H| \to \infty$ as $N \to \infty$. On the other hand, whenever $\hat{\boldsymbol{\beta}}_H$ converges to a real number $\boldsymbol{\beta}$ (say), in view of the continuity of Δ, $\Delta(\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\beta}}_H)$ must converge to $\Delta(\boldsymbol{\alpha}^*, \boldsymbol{\beta})$ on that set of probability one. Since any bounded sequence must have a convergent subsequence, it is now obvious that $\hat{\boldsymbol{\beta}}_H$ must converge to $\boldsymbol{\beta}^*$, where $\Delta(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \Delta(\boldsymbol{\alpha}, \boldsymbol{\beta})$, which is same as the Bayes risk in this case.

For prior probabilities π_1 and π_2 (π_1 not necessarily equal to π_2), and for two

competing normally distributed populations with parameters (μ_1, Σ) and (μ_2, Σ),

$$\begin{aligned}
&\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x}) \\
&\quad \Leftrightarrow \pi_1 \, |\Sigma|^{-1/2} \, e^{-\frac{1}{2}(\mathbf{x}-\mu_1)' \Sigma^{-1}(\mathbf{x}-\mu_1)} > \pi_2 \, |\Sigma|^{-1/2} \, e^{-\frac{1}{2}(\mathbf{x}-\mu_2)' \Sigma^{-1}(\mathbf{x}-\mu_2)} \\
&\quad \Leftrightarrow (\mathbf{x}-\mu_2)' \Sigma^{-1}(\mathbf{x}-\mu_2) - (\mathbf{x}-\mu_1)' \Sigma^{-1}(\mathbf{x}-\mu_1) > C, \quad \text{where } C = 2 \, log \, (\pi_2/\pi_1). \\
&\quad \Leftrightarrow 2\mathbf{x}' \Sigma^{-1}(\mu_1 - \mu_2) > \{\mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2\} + C.
\end{aligned}$$

Therefore, the optimum Bayes rule is indeed unique, and it is linear in nature. Finally, as U and Δ are both continuous functions in this case of multivariate normal distribution, the proof of the corollary is complete.

Proof of Corollary 5.3 : It suffices to show that under the given conditions, the optimum quadratic classifier is the unique Bayes classifier. When the two competing population distributions are multivariate normal with location and scatter parameters (μ_1, Σ_1) and (μ_2, Σ_2),

$$\begin{aligned}
&\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x}) \\
&\quad \Leftrightarrow \pi_1 \, |\Sigma_1|^{-1/2} \, e^{-\frac{1}{2}(\mathbf{x}-\mu_1)' \Sigma_1^{-1}(\mathbf{x}-\mu_1)} > \pi_2 \, |\Sigma_2|^{-1/2} \, e^{-\frac{1}{2}(\mathbf{x}-\mu_2)' \Sigma_2^{-1}(\mathbf{x}-\mu_2)} \\
&\quad \Leftrightarrow (\mathbf{x}-\mu_2)' \Sigma_2^{-1}(\mathbf{x}-\mu_2) - (\mathbf{x}-\mu_1)' \Sigma_1^{-1}(\mathbf{x}-\mu_1) > C,
\end{aligned}$$

where $C = 2 \, log \left(\frac{\pi_2 |\Sigma_1|^{1/2}}{\pi_1 |\Sigma_2|^{1/2}} \right)$. Therefore, the optimum Bayes rule is indeed unique and quadratic in nature.

Probability density function $f(\mathbf{x})$ of a d-dimensional elliptically symmetric Pearson type VII distribution is given by

$$f(\mathbf{x}) = C_d \, |\Sigma|^{-1/2} \{1 + \nu^{-1}(\mathbf{x}-\mu)' \Sigma^{-1}(\mathbf{x}-\mu)\}^{-M},$$

where μ and Σ are the location and scatter parameters, $\nu > 0$, $M > d/2$ and $C_d = (\pi\nu)^{-d/2}\Gamma(M)/\Gamma(M - d/2)$. Now, consider two Pearson type VII distributions, which are of the same form except possibly for their location and scatter parameters. Let μ_i and Σ_i be the location parameter and the scatter matrix for the i-th (i=1,2) population, and π_i be its prior probability. Then, $\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x})$

$$\begin{aligned}
&\Leftrightarrow \pi_1 |\Sigma_1|^{-1/2} \{1 + \nu^{-1}(\mathbf{x}-\mu_1)' \Sigma_1^{-1}(\mathbf{x}-\mu_1)\}^{-M} > \pi_2 |\Sigma_2|^{-1/2} \{1 + \nu^{-1}(\mathbf{x}-\mu_2)' \Sigma_2^{-1}(\mathbf{x}-\mu_2)\}^{-M} \\
&\Leftrightarrow \left\{ \frac{1 + \nu^{-1}(\mathbf{x}-\mu_1)' \Sigma_1^{-1}(\mathbf{x}-\mu_1)}{1 + \nu^{-1}(\mathbf{x}-\mu_2)' \Sigma_2^{-1}(\mathbf{x}-\mu_2)} \right\}^{-M} > m \quad \text{for } m = \frac{\pi_2 |\Sigma_2|^{-1/2}}{\pi_1 |\Sigma_1|^{-1/2}} \\
&\Leftrightarrow \left\{ \frac{\nu + (\mathbf{x}-\mu_1)' \Sigma_1^{-1}(\mathbf{x}-\mu_1)}{\nu + (\mathbf{x}-\mu_2)' \Sigma_2^{-1}(\mathbf{x}-\mu_2)} \right\} < m_0 = m^{-1/M} \\
&\Leftrightarrow (\mathbf{x}-\mu_1)' \Sigma_1^{-1}(\mathbf{x}-\mu_1) - m_0(\mathbf{x}-\mu_2)' \Sigma_2^{-1}(\mathbf{x}-\mu_2) - (m_0 - 1)\nu < 0.
\end{aligned}$$

Clearly, the left hand side of last inequality above is a quadratic function of $\mathbf{x}$. Therefore once again the optimum Bayes rule is unique, and it turns out to be a quadratic classifier.

Chapter 6

Maximum depth classifiers

6.1 Main problem and motivation

While in the preceding chapter, we have developed and investigated some semiparametric classification procedures, the objective in this chapter is to demonstrate that various notions of data depth can also be used to develop fully nonparametric classifiers. Note that, data depth measures the centrality of a d-dimensional observation $\mathbf{x}$ with respect to a multivariate distribution F or with respect to a given d-dimensional data cloud. It helps to build up a systematic and nonparametric approach to generalize the graphical and distributional properties of univariate distributions to multivariate distributions. The notions of multivariate median, multivariate L-statistics, tests for the center of elliptical symmetry, measures of multivariate dispersion and skewness are some of the well known examples of its application (see e.g., Hodges, 1955; Chaudhuri and Sengupta, 1993; Liu and Singh, 1993; Liu, Parelius and Singh, 1999; Vardi and Zhang, 2000; Mosler, 2002). Half-space depth (HD) (see e.g., Tukey, 1975), simplicial depth (SD) (see e.g., Liu ,1990), majority depth (MJD) (see e.g., Singh, 1991; Liu and Singh, 1993), Mahalanobis depth (MD) (see e.g., Mahalanobis, 1936, Liu and Singh, 1993) and projection depth (PD) (see e.g., Stahel, 1981; Donoho, 1982) are some of the popular depth functions available in the literature. Various other well known depth functions like likelihood depth, convex hull peeling depth and zonoid depth have been studied by Koshevoy and Mosler (1997), Liu, Parelius and Singh (1999), Zuo and Serfling (2000a, 2000b) and Mosler (2002).

Zuo and Serfling introduced the notion of simplicial volume depth (SVD), which is related to Oja median (Oja, 1983). SVD of an observation $\mathbf{x}$ with respect to the distribution F can be expressed as

$$SVD^{\delta}(F,\mathbf{x}) = \left\{1 + E_F\left[\frac{\nabla\{\mathbf{x},\mathbf{X}_1,\ldots,\mathbf{X}_d\}}{|\mathbf{\Sigma}_F|^{1/2}}\right]^{\delta}\right\}^{-1},$$

where $X_1, \ldots, X_d$ are observations from F, $\nabla\{\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_d\}$ is the volume of the d-dimensional simplex formed by $\mathbf{x}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_d$, and Σ_F is the scatter matrix of the distribution F. Note that, the division by $|\Sigma_F|^{1/2}$ is required only to make the depth function affine invariant.

Recently, following the work of Chaudhuri (1996) and Kolchinskii (1997) on spatial quantiles, Vardi and Zhang (2000) and Serfling (2002) came up with another notion of location depth known as spatial depth (*SPD*). Jornsten, Vardi and Zhang (2002) and Jornsten (2004) used this depth function to develop robust clustering and classification techniques. Spatial depth of an observation $\mathbf{x}$ with respect to the distribution F is defined to be $SPD(F, \mathbf{x}) = 1 - \|E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}\|$, where $\mathbf{X} \sim F$. This depth function has some nice theoretical properties. When $d \geq 2$, for all F, $E_F\{(\mathbf{x}-\mathbf{X})/\|\mathbf{x}-\mathbf{X}\|\}$ is continuous and monotonic transformation on R^d, and it uniquely determines the distribution function $F(\mathbf{x})$ (see e.g., Koltchinskii, 1997). When the observation $\mathbf{x}$ is located near the center of the distribution, $E_F\{(\mathbf{x}-\mathbf{X})/\|\mathbf{x}-\mathbf{X}\|\}$ is expected to be very close to 0, and hence $SPD(F, \mathbf{x})$ is expected to attain its maximum value 1. On the other hand, if we move away from the center, SPD approaches to 0. Unlike the other depth functions, SPD is easy to compute for high dimensional data, and one can define SPD for infinite dimensional Hilbert spaces (see e.g., Chaudhuri, 1996) as well.

Sample versions of various depth functions are obtained by replacing F with the empirical distribution function F_n that puts mass $1/n$ on each of the n data points in d-dimensional space. Theoretical properties of these empirical depths and their corresponding depth contours have been extensively studied (see e.g. Liu, 1990; Nolan, 1992; Donoho & Gasko, 1992; Liu & Singh, 1993; He and Wang, 1997; Zuo and Serfling, 2000a, 2000b, Mizera, 2002; Mizera and Volauf, 2002) in the literature. To make it notationally simpler, instead of $D(F_j, \mathbf{x})$ and $D(F_{n_j}, \mathbf{x})$ we use $D(j, \mathbf{x})$ and $D_n(j, \mathbf{x})$, respectively, to denote the theoretical and the empirical depth of $\mathbf{x}$ in the j^{th} population. *Throughout this chapter, all the population distributions are assumed to possess densities which are continuous and positive over the entire d-dimensional space.*

Unlike the depth based semiparametric classifiers investigated in Chapter 5, the maximum depth classifiers studied in this chapter do not assume any specific parametric form of the separating surface, and they classify an observation to the class with respect to which it has the maximum location depth. Like nearest neighbor classification (see Chapter 4), these classifiers do not require any learning from the training sample, and the full training data have to be stored to classify the future observations. These maximum depth classifiers can be expressed as

$$\mathbf{d}_D(\mathbf{x}) = \arg\max_j \; D_{n_j}(j, \mathbf{x}),$$

where n_j is the number of training sample observations and $D_{n_j}(j, \mathbf{x})$ is the empirical depth of $\mathbf{x}$ in the j-th population. When the prior probabilities of the competing classes are all equal, one can use various notions of depths to construct different maximum depth classifiers for discriminating among several competing populations. Note that when the competing populations have the same scatter matrix (e.g., if the population distributions satisfy a location shift model), it is not necessary to have $|\Sigma|^{1/2}$ in the denominator of the expression SVD^{δ} when it is used for maximum depth classification.

6.2 Misclassification rates and asymptotic optimality

When the population distribution is elliptic with density function strictly decreasing in every directions from its center of symmetry, some of the depth functions also hold that monotonicity property (see e.g., Zuo and Serfling, 2000a), and they turn out to be a decreasing function of the Mahalanobis distance (see Mahalanobis, 1936). HD, SD, MJD, MD, PD and SVD (for $\delta \geq 1$) are some of the depth functions which belong to this family (see e.g. Liu, 1990; Singh, 1991; Donoho and Gasko, 1992; Nolan, 1992; Liu & Singh, 1993, Zuo and Serfling, 2000a). Therefore, in equal prior cases and when several elliptic populations differ only in their location parameters, these depth functions are equivalent to Mahalanobis distance for classification, and they lead to the optimal Bayes classifier. MD is the simplest one to calculate, and it leads to the usual linear discriminant function based on the first and the second moments of the training sample observations. Like MD, SVD also depends on the moments of the training sample and consequently they both are sensitive to outlier and extreme values. Many other classifiers derived from different depth functions are not based on moments, and they are more suitable when the training set observations have distributions with heavy tails.

SPD in practice has some advantages over the other depth measures. We have already pointed out that it is computationally less expensive than most of the depth functions, and can be used for classification in infinite dimensional Hilbert spaces. Since empirical version of SPD is continuous in $\mathbf{x}$, there is almost no possibility of ties which may be a problem for depth functions like HD, SD and MJD because of their step function (piecewise constant) like nature. SPD is also invariant under orthogonal and scale transformation. Therefore, in the case of spherically symmetric distributions, where the density function decreases with the distance from the center, SPD also holds that monotonicity property (see Proposition 6.1 in Section 6.8). We now state the following theorems on the asymptotic optimality of different maximum depth classifiers.

Theorem 6.1 : *Suppose that the population density functions $f_1, f_2, \ldots, f_J$ are elliptically symmetric, and $f_j(\mathbf{x}) = g(\mathbf{x}-\mu_j)$ for some location parameters μ_j and a common*

density function g with $g(k\mathbf{x}) \leq g(\mathbf{x})$ for every $\mathbf{x}$ and $k > 1$. Now define $\mathbf{n} = (n_1, n_2, \ldots, n_J)$ as the vector of training sample sizes for different classes, and $\Delta_{\mathbf{n}}$ as the error rate of the corresponding empirical depth based classifier. Then, in the equal prior cases, for HD, SD, MJD and PD, $\Delta_{\mathbf{n}}$ converges to the optimal Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.

Theorem 6.2 : *Assume the same set up as in Theorem 6.1. If g is spherical, $\Delta_{\mathbf{n}}$ in the case of SPD converges to the optimal Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.*

Theorem 6.3 : *Assume all the conditions of Theorem 6.1. For some given $\mathbf{x}$, define $\nabla_j\{\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_d\}$ as the volume of the d-dimensional simplex formed by $\mathbf{x}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_d$, which are observations from f_j. Further assume that $E_{f_j}[\nabla_j\{\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_d\}]^\delta < \infty$ for all $j = 1, 2, \ldots, J$, and some $\delta \geq 1$. Then, $\Delta_{\mathbf{n}}$ in the case of SVD^δ converges to optimal Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.*

6.3 Data analytic implementation of the classifiers

Due to difficulty in computing depths in higher dimension, for all data analytic purposes we restrict ourselves to HD, SD and SPD only. SPD is the easiest one to compute but given a data cloud of n observations, the computational cost for HD and SD of an observation increases rapidly with the dimension at a geometric rate (see e.g. Chaudhuri and Sengupta, 1993; Rouseeuw and Ruts, 1996; Rousseeuw and Struyf, 1998). Therefore, exact computation of these depths is not feasible for high dimensional problems, and there one can only use some approximate version. Such an approximate version for HD was proposed in Section 5.4. This approximation allows us to use derivatives of certain smooth functions to find out the direction of steepest ascent/descent of the objective function to be optimized. In this chapter, for all problems with $d > 2$, we have adopted the same idea for computing HD of an observation. Exact version of HD is used for bivariate data only. In order to cope up with the problem of possible presence of several local optima, we have always run our approximate version of the optimization algorithm a few times starting from different random initial points. Since no such approximate algorithm is available for SD, we have used this depth function only for two dimensional problems.

Apart from computational difficulty, HD and SD have another problem in higher dimension due to sparsity of the data. Consider the following example on a two-class problem where the classes are $N_2(0, 0, 1, 1, 0)$ and $N_2(2, 2, 1, 1, 0)$. We have generated 100 observations from each class, and the scatter plot of this data set is given in Figure 6.1.

In this figure, one can notice some observations which have zero depth for both the populations. For instance, the observations 'A' and 'B' (see Figure 6.1) clearly belong to two different classes but both have zero depth for both of the two populations. Clearly,

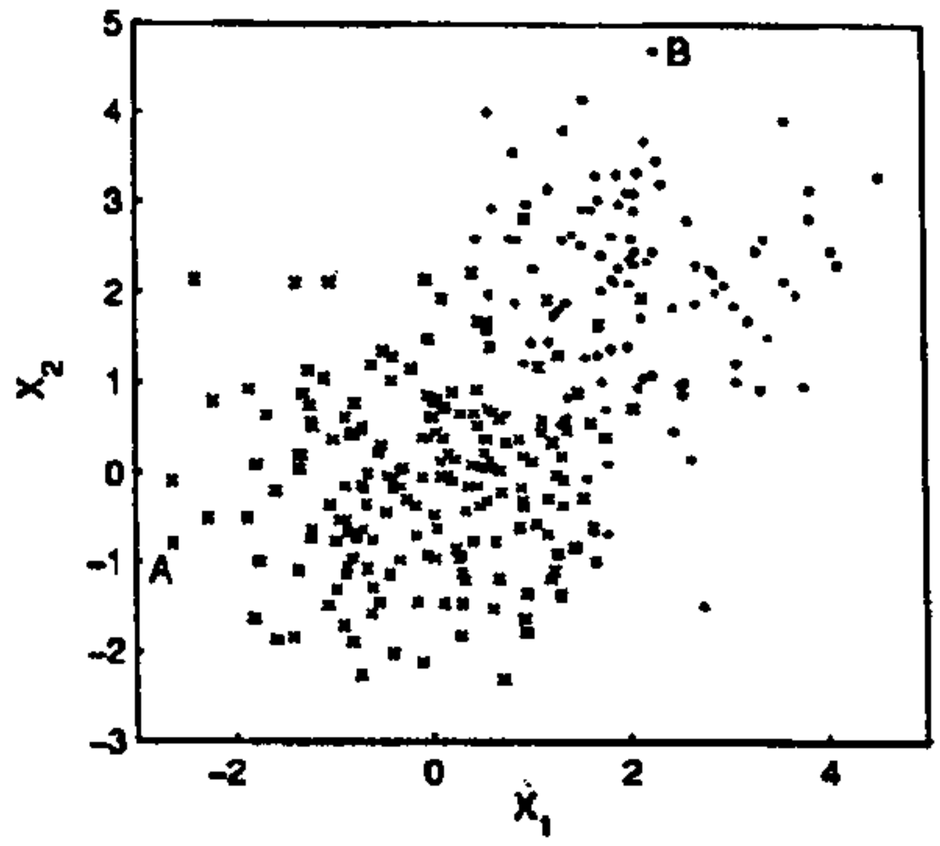


Figure 6.1 : Scatter plot for simulated data

the classifier based on HD and SD fail to classify these two observations correctly. In high dimensions, when the sizes of the training samples are small compared to the dimensionality of the problem, we have a high proportion of observations having zero depth for both of the competing populations. For these observations, instead of using HD or SD, we have used the 1-nearest neighbor (1-NN) rule (see Chapter 4) to classify them. However, from the definition of SPD it is quite transparent that apart from one-dimensional cases, when the distribution is not supported on a real line, SPD is always positive. This differs critically from HD and SD, which attain the value zero at points outside the support of the distribution, when the support is bounded. Because of this property, unlike HD and SD, SPD does not get affected by data sparsity in high dimensions.

6.4 Numerical results

In this section, we use some simulated and some benchmark data sets to illustrate the performance of the maximum depth classifiers. Performance of traditional LDA and QDA on those data sets has also been given to facilitate the comparison. In the case of simulated examples, we report the corresponding Bayes errors as well. For benchmark data sets, where optimal errors are not known, we reported the misclassification rates of some nonparametric methods to compare the performance of the maximum depth classifiers (d_D). *In all these data sets, for computing SPD of an observation, we standardized the measurements variables in each class using inter-quartile ranges.* This enables us to use SPD even if different measurement variables are originally not of comparable units. Throughout this section, prior probabilities for all competing populations are considered to be equal.

6.4.1 Results on simulated data sets

As simulated examples, we consider some two-class problems, where both the populations are elliptically symmetric (multivariate normal and multivariate Cauchy) and they differ only in their location parameters. To make our examples simpler, we consider $\mu_1 = (0,\ldots,0)$ and $\mu_2 = (\mu,\ldots,\mu)$ as the location parameters of the two classes while **I** is taken as the common scatter matrix for the two populations. We use μ =1 and μ=2 for our experiment. For each of these simulated examples, we generate a training set taking equal number (100 or 200) of observations from each class, while a test set of 1000 observations (500 from each class) is used to compute the misclassification rates for different classifiers. Each experiment is carried out 100 times, and the average misclassification rates and their corresponding standard errors over those 100 simulations are reported in Table 6.1 and 6.2. For two dimensional problems, we report the performance of HD classifier based on its exact version, whereas the approximate version is used for dimensions larger than two. Due to computational difficulty, SD is used only in the case of bivariate problems.

Distribution	μ	Bayes risk	n	LDA	QDA	SPD	HD	SD
Normal	1	23.98	100	24.12(0.15)	24.23(0.15)	24.45(0.16)	24.94(0.17)	25.05(0.17)
			200	24.16(0.13)	24.19(0.13)	24.29(0.12)	24.64(0.13)	24.71(0.13)
	2	7.87	100	8.05(0.09)	8.07(0.09)	8.21(0.09)	8.44(0.10)	8.64(0.11)
			200	7.92(0.09)	7.95(0.08)	8.01(0.09)	8.16(0.09)	8.25(0.09)
Cauchy	1	30.40	100	41.99(0.81)	49.67(0.14)	32.81(0.20)	33.48(0.27)	33.57(0.25)
			200	43.26(0.96)	49.80(0.19)	32.25(0.19)	32.89(0.21)	32.97(0.22)
	2	19.58	100	33.05(1.29)	47.81(0.54)	21.84(0.20)	22.65(0.23)	22.75(0.23)
			200	34.42(1.41)	49.37(0.18)	21.05(0.16)	21.91(0.20)	22.00(0.19)

Table 6.1 : Misclassification rates (in %) on elliptic distributions (dimension 2)

Distribution	μ	Bayes risk	n	LDA	QDA	SPD	HD
Normal	1	19.32	100	19.63(0.13)	19.83(0.13)	20.01(0.14)	21.27 (0.13)
			200	19.60(0.11)	19.78(0.11)	19.85(0.11)	20.52(0.13)
	2	4.16	100	4.28(0.07)	4.34(0.07)	4.46(0.07)	5.09(0.09)
			200	4.20(0.07)	4.26(0.07)	4.33(0.07)	4.68(0.08)
Cauchy	1	27.29	100	39.78(0.77)	49.77(0.10)	31.09(0.26)	32.69(0.31)
			200	39.41(0.94)	49.78(0.13)	29.65(0.21)	31.14(0.26)
	2	16.67	100	27.51(1.12)	46.13(0.77)	19.51(0.20)	21.33(0.27)
			200	27.09(1.04)	48.47(0.46)	18.68(0.17)	20.02(0.18)

Table 6.2 : Misclassification rates (in %) on elliptic distributions (dimension 3)

In the case of normal distributions, as expected, LDA led to the best performance, and it could nearly achieve the optimal Bayes risk. Error rates for QDA were also quite comparable. The maximum depth classifiers could produce satisfactory performance as well. SPD had a slight edge over the other two depth based classifiers. However, the performance of LDA and QDA falls drastically when the observations are generated from heavy tailed

distributions. In the case of Cauchy distributions, HD, SD and SPD clearly outperformed the traditional methods, and their error rates were very close to the optimal Bayes risks.

In all these simulated examples, even when 100 observations are taken from each class, the maximum depth classifiers could achieve error rates fairly close to the optimal Bayes risk, which became even closer for larger sample sizes. The following theorem gives some idea about the asymptotic accuracy of misclassification rates for some empirical depth based classifiers under some regularity conditions.

Theorem 6.4 : *Suppose that the population density functions $f_1, f_2, \ldots, f_J$ satisfy the conditions of Theorem 6.1 and define $\Delta_{\mathbf{n}}$ as before. Also define $D^{0j}(\mathbf{x}) = \min_{\{i\,:\,i \neq j\}} \{D(j,\mathbf{x}) - D(i,\mathbf{x})\}$ and Δ = optimal Bayes risk. Then, in the equal prior cases, we have*

$$\Delta_{\mathbf{n}} < \Delta + \frac{1}{J}\sum_{j=1}^{J} \int_{D^{0j}(\mathbf{x})>0} [1 - \beta_{\mathbf{n}}\{D^{0j}(\mathbf{x})\}] f_j(\mathbf{x}) d\mathbf{x}$$

for some appropriate function $\beta_{\mathbf{n}}$, which depends on the choice of the depth measure, but for any t $(0 < t < \infty)$, $\beta_{\mathbf{n}}(t) \to 1$ as $min\{n_1, n_2, \ldots, n_J\} \to \infty$. Here, for HD and SD, $\beta_{\mathbf{n}}(t)$ is of the form $\beta_{\mathbf{n}}(t) = \prod_j max\{0, 1 - 2n_j^d e^{-n_j t^2/2}\}$ and $\beta_{\mathbf{n}}(t) = \prod_j max\{0, 1 - 2e^{-\lfloor n_j/d+1 \rfloor t^2/2}\}$, respectively, where $\lfloor x \rfloor$ denotes the highest integer $\leq x$. Further, if the population distributions are spherical, error rate of the SPD classifier also satisfies the above inequality with $\beta_{\mathbf{n}}(t) = \prod_{j=1}^{J} max\{0, 1 - 2d\ e^{-n_j t^4/8d^2}\}$.

6.4.2 Results on "synthetic data"

Our next illustration is using the "synthetic data", which has already been discussed several times in previous chapters. It is a well-known benchmark data set, where both the classes are equal mixtures of two Gaussian distributions. Since the two populations have equal number of observations both in training and test sets, the prior probabilities for these two populations are taken to be equal. It is known that the optimal Bayes risk for this data set is 8.0%. On this data set, QDA led to best error rate of 10.2%. Error rate for LDA (10.8%) and that of the SPD classifier (10.5%) were also fairly competitive. However, for HD (error rate = 12.8%) and SD (error rate = 13.8%), we had slightly higher error rates.

6.4.3 Results from the analysis of real data sets

We used two other benchmark data sets, namely vowel recognition data-1 and salmon data, for further illustration. Like synthetic data, salmon data also have the same number of observations from two different populations - which justifies the use of equal priors for the competing classes. Since the sample sizes of the different classes in vowel data-1 are not

very different, we have taken the priors to be equal for evaluating the performance of the depth based classification techniques.

On vowel recognition data-1, QDA could achieve to an error rate of 19.8%. The usual k-nearest neighbor method based on Euclidean distance and cross validated estimate of k also had reasonably lower misclassification rate (21.9%). HD classifier and the classification tree method both led to the same error rate of 23.7%. Misclassification rates for LDA (25.2%) and SD classifier (32.7%) were higher than that of the other classification methods. SPD classifier led to an error rate of 24.6%. However, when the data points were standardized using an estimate of pooled dispersion matrix, this misclassification rate reduced to 21.3%. On this standardized data set, classification tree and nearest neighbor method had error rates of 24.0% and 19.2%, respectively. Since LDA, QDA and the classifiers based on HD and SD are affine invariant, the result remains same whether one uses standardized or unstandardized data.

Unlike the other two data sets, salmon data does not have separate training and test sets. We divided this data set randomly to form the training sets consisting of 80 observations (40 from each class), while the remaining 20 observations were used to form the corresponding test sets. This random division was carried out 250 times. The average misclassification rates for different methods and the corresponding standard errors over these 250 partitions are reported here. On this data set, QDA led to the best error rate of 7.23% with a standard error of 0.32%. LDA (error rate =7.54%, S.E.=0.32%), SPD (error rate =7.46%, S.E.=0.33%) and HD (error rate =7.32%, S.E.=0.34%) classifiers could achieve fairly competitive performance. SD classifier and k-nearest neighbor method failed to compete with the above classifiers. For these two cases, the misclassification rates were found to be 8.76% and 8.02% respectively with standard errors of 0.41% and 0.36% in the respective cases. For this data set also, we applied the nearest neighbor and SPD classifier after standardization by pooled dispersion matrix. On this standardized version of the data, the nearest neighbor method and the SPD classifier could achieve average test set error rates of 8.11% (S.E. = 0.37%) and 7.42% (S.E. = 0.34%), respectively. Due to computational difficulties in finding the error rates over repeated partitions, we could not report the performance of the classification tree method for this data set.

From the analysis of these data sets, SPD classifier appears to be the best among the maximum depth classifiers. Both in terms of computational cost and misclassification rate, it had a clear edge over HD and SD classifiers. Its performance on the simulated and benchmark data sets was fairly competitive as compared to the other parametric and nonparametric classification procedures.

6.5 Classification problems with unequal priors

In Sections 6.2 and 6.4, we have already observed that in the equal prior case and under the location shift model, depth based classification methods can be used as good nonparametric distribution-free alternatives for the traditional methods of discriminant analysis. However, in practice, different populations may have different priors and they may not belong to the same family of elliptic distributions. In this section, we deal with such situations and propose another depth based classifier which is capable of achieving reasonably lower misclassification rates under a more general set up that incudes unequal prior cases.

Theorem 6.5 : *When the population distributions are elliptically symmetric, for MD, HD, SD, MJD, SVD and PD, there exist some functions $\theta_j(\cdot)$ of population depth $D(j,\mathbf{x})$ (θ_j may depend on the type of the depth function) such that the optimal Bayes classifier is given by*

$$\mathrm{d}_B(\mathbf{x}) = arg\max_j \pi_j \theta_j\{D(j,\mathbf{x})\},$$

where the π_j's are the prior probabilities of different classes.

Note that when the population distributions satisfy a location shift model, and the density functions decrease with the Mahalanobis distance from the center of symmetry, the functions θ_j's are same for all the populations, and they are monotonic in nature. Therefore, in the equal prior cases and under the above conditions, this Bayes classifier turns out to be the maximum depth classifier. To construct a classification rule based on the training sample observations, one needs to find out appropriate sample analogs for $\theta_j\{D(j,\mathbf{x})\}$. However, in practice, for most of the depth functions, $\theta_j\{D(j,\mathbf{x})\}$ is a complicated function of $D(j,\mathbf{x})$, and it is very difficult to obtain its consistent estimate based on training sample observations. Of course, because of the simple relation between Mahalanobis distance and HD (see the proof of Lemma 6.1 in Section 6.8) in the case of elliptic distributions (with location parameter $\boldsymbol{\mu}_j$ and scatter parameter $\boldsymbol{\Sigma}_j$), it is possible to have a simpler expression for $\theta_j\{D(j,\mathbf{x})\}$ when HD is used. In that case, $\theta_j\{D(j,\mathbf{x})\}$ can be expressed as

$$\theta_j\{D(j,\mathbf{x})\} = |\boldsymbol{\Sigma}_j|^{-1/2}\ \varrho_j\left(\gamma_j\{D(j,\mathbf{x})\}\right) / \left(\gamma_j\{D(j,\mathbf{x})\}\right)^{d-1},$$

where the depth function $D(j,\mathbf{x})$ and the Mahalanobis distance $\gamma_j\{D(j,\mathbf{x})\}$ have the relation $D(j,\mathbf{x}) = 1 - F_j\left(\gamma_j\{D(j,\mathbf{x})\}\right)$ for $F_j(\cdot)$ being the distribution function of $\alpha' \boldsymbol{\Sigma}_j^{-1}(\mathrm{X}_j - \boldsymbol{\mu}_j)$ for any α with $\|\alpha\| = 1$, and $\varrho_j(\cdot)$ is the density function of $\gamma_j\{D(j,\mathbf{x})\}$. Consistent estimates of $\boldsymbol{\Sigma}_j$, $\gamma_j\{D(j,\mathbf{x})\}$ and its density function $\varrho_j(\cdot)$ lead to a decision rule capable of achieving misclassification rates close to the optimal Bayes risk. Instead of plugging any moment based estimates of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, here we use half-space depth to estimate γ_j while kernel density estimation technique is used to estimate ϱ_j and to construct the modified depth classifier (see the following section for detailed description). It requires a consistent

estimate for $|\Sigma_j|$ as well. To make the classification procedure stable against outliers and contaminated observations, one has to plug in some robust estimate for Σ_j. One can also bypass this estimation problem by writing the classifier in the form

$$\mathbf{d}(\mathbf{x}) = \arg\max_j C_j \; \varrho_j \left(\gamma_j\{D(j,\mathbf{x})\}\right) / \left(\gamma_j\{D(j,\mathbf{x})\}\right)^{d-1},$$

where $C_1, C_2, \ldots, C_J$ are suitable constants. Clearly, the error rate of this classifier depends on these constants, and by minimizing this error rate with respect $C_1, C_2, \ldots, C_J$, one arrives at the optimal Bayes rule. In practice, one can take $C_1 = 1$ and minimize the error rate over the other $J-1$ independent parameters. Similarly, after finding the consistent estimates for γ_j and ϱ_j, using the training sample, one can minimize the error rate of the resulting classifier to build up the final classification rule.

6.5.1 Description of the methodology and related convergence properties

Let us start with a two-class problem, where we have observations $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ from the j^{th} population g_j $(j = 1, 2)$, and we want to classify a future observation $\mathbf{x}_0$ into one of these two classes. At first, we compute empirical half-space depths $D_{n_j}(j, \mathbf{x}_0)$ of $\mathbf{x}_0$ with respect to the data cloud of the two populations $(j = 1, 2)$. Next, we project the observations of g_j in some fixed direction $\boldsymbol{\alpha}$ $(\|\boldsymbol{\alpha}\| = 1)$ and find out two points a_1 and a_2 such that they both have empirical depth $D_{n_j}(j, \mathbf{x}_0)$ but lie on the opposite side of the center. Half of the distance between these two points (i.e. $|a_1 - a_2|/2$) is taken as an estimate for the re-scaled Mahalanobis distance $\gamma_j\{D(j, \mathbf{x}_0)\}\sqrt{\boldsymbol{\alpha}'\Sigma_j\boldsymbol{\alpha}}$. The following theorem establishes the consistency of this estimate under appropriate condition.

Theorem 6.6 : *Suppose that* $\mathbf{X}$ *has an elliptically symmetric density with location parameter* μ *and scatter matrix* Σ. *Let* δ_n *be the empirical depth of an observation* $\mathbf{x}$ *with respect to a data cloud of* n *i.i.d. observations from the same distribution that of* $\mathbf{X}$. *Also define,* $\xi_{p,\boldsymbol{\alpha},n}$ *as the* p*-th empirical quantile of* $\boldsymbol{\alpha}'\mathbf{X}$ *for some* $\boldsymbol{\alpha}$ *with* $\|\boldsymbol{\alpha}\| = 1$. *Then, as* $n \to \infty$, $(\xi_{1-\delta_n,\boldsymbol{\alpha},n} - \xi_{\delta_n,\boldsymbol{\alpha},n})/2 \xrightarrow{a.s} \{(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\}^{1/2}\sqrt{\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}}$.

This estimation procedure is repeated using a number of different directions $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, ..., $\boldsymbol{\alpha}_S$ and the average of these estimates is taken as the final estimate $\hat{v}_0^{(j)}$ for the re-scaled Mahalanobis distance $v_0^{(j)} = \lambda_j \; \gamma_j\{D(j, \mathbf{x}_0)\}$, where $\lambda_j = \sum_{s=1}^{S} \sqrt{\boldsymbol{\alpha}_s'\Sigma_j\boldsymbol{\alpha}_s}/S$. It should also be noted that the form of the modified depth classifier remains invariant under such scale transformation and only the constant terms C_1, C_2 (as described in the previous section) get changed to C_1^*, C_2^*, where $C_2^*/C_1^* = (\lambda_1/\lambda_2)^d C_2/C_1$. Not only for $\mathbf{x}_0$, we estimate the re-scaled Mahalanobis distance at each data point using leave-one-out (leaving out that particular data point) method. In this way, a number of bivariate observations $\hat{\mathbf{v}}_{11}, \hat{\mathbf{v}}_{12}, \ldots, \hat{\mathbf{v}}_{1n_1}$ and $\hat{\mathbf{v}}_{21}, \hat{\mathbf{v}}_{22}, \ldots, \hat{\mathbf{v}}_{2n_2}$ is obtained, where $\hat{\mathbf{v}}_{ji} = (\hat{v}_{ji}^{(1)}, \hat{v}_{ji}^{(2)})$ denotes the

estimate of $(v_{ij}^{(1)}, v_{ji}^{(2)})$, the re-scaled Mahalanobis distances of $\mathbf{x}_{ji}$ form the centers of the first and the second populations. Using $\hat{v}_{ji}^{(j)}$, $i = 1, 2, \ldots, n_j$ as the observations from the j-th population ($j = 1, 2$), we estimate the density function (Υ_j, say) of the re-scaled Mahalanobis distance by a kernel method (see e.g., Silverman, 1986; Scott, 1992) [note that $\Upsilon_j(v) = \lambda_j \, \varrho_j(\lambda_j \, v)$]. Instead of using bandwidths that target to minimize the estimated mean integrated square error of a kernel density estimate, for classification problems it is better to use the largest bandwidth that minimizes the cross-validated misclassification rate (see Section 2.4 and 2.5 in Chapter 2). However, To find out the cross-validated error rates, one has to estimate the value of $C^* = C_2^*/C_1^*$ as well. Here, we use leave-one-out cross-validation technique for simultaneous estimation C^* and the bandwidths. Let h_{1n_1} and h_{2n_2} be the estimated bandwidths, and $\hat{\Upsilon}^*_{1h_{1n_1}}$ and $\hat{\Upsilon}^*_{2h_{2n_2}}$ be the kernel estimates of the densities of the re-scaled Mahalanobis distances for the two populations. Now, we classify the future observation $\mathbf{x}_0$ to population-1 if and only if $\hat{\Upsilon}^*_{1h_{1n_1}}(\hat{v}_0^{(1)})/\{\hat{v}_0^{(1)}\}^{d-1} > \hat{C}^* \hat{\Upsilon}^*_{2h_{2n_2}}(\hat{v}_0^{(2)})/\{\hat{v}_0^{(2)}\}^{d-1}$, where $\hat{C}^*$ is the estimate of C^* obtained by cross-validation. From Theorem 6.6, it is quite transparent that for $j = 1, 2$, $\hat{v}_0^{(j)}$ converges almost surely to $v_0^{(j)}$. Again, under some appropriate regularity conditions (see Proposition 6.2 in Section 6.8) $\hat{\Upsilon}^*_{jh_{jn_j}}(\hat{v}_0^{(j)})$ converges to $\Upsilon_j(v_0)$ as well. Therefore, suitable estimates of C^* should lead to misclassification rates close to the optimal Bayes risk.

For classification problems with more than two-populations, we adopt similar strategy to find out $\hat{v}_0^{(j)}$ and $\hat{\Upsilon}^*_{jh_{jn_j}}(\hat{v}_0^{(j)})$ for $j = 1, 2, \ldots, J$. Then, one can construct a modified depth classifier of the form

$$\mathrm{d}_{D^*}(\mathbf{x}_0) = \arg\max_j C_j \; \hat{\Upsilon}^*_{jh_{jn_j}}(\hat{v}_0^{(j)})/\{\hat{v}_0^{(j)}\}^{d-1}$$

The error rate of this classifier depends on $J-1$ independent parameters $C_2/C_1, C_3/C_1$, ..., C_J/C_1 and minimizing this error rate over those parameters the final classifier can be obtained.

Unlike the maximum depth classifier, the performance of this classifier does not get affected by deviation from location shift model or violation of monotonic nature of the density functions, and it works for general case, where the prior probabilities may or may not be equal. However, it is computationally difficult to minimize the error rate simultaneously with respect to $C_1, C_2, \ldots, C_J$ as well as $h_1, h_2, \ldots, h_J$. Instead, one may split the multi-class problem into a number of two-class problems taking each pair of classes at a time and proceed in the same way as before. Then, results of these pairwise comparisons may be combined by the voting method (see e.g., Friedman, 1996) to arrive at the final classification.

6.6 Numerical results for unequal prior cases

In this section, we use some simulated and benchmark where the competing populations have different prior probabilities. Along with the results of the proposed modified half-space depth classifier (henceforth called HD^*), the performance of LDA and QDA are also reported for proper evaluation of this classification methodology. For simulated examples, Bayes errors are reported as well. Performance of the nearest neighbor classifier is given for the biomedical data set to facilitate comparison.

6.6.1 Results on simulated data sets

For simulation experiments, we consider two dimensional problems only. We choose the same two-dimensional examples on normal and Cauchy populations as discussed in Section 6.4, where the competing populations satisfy a location shift model. But this time we use unequal priors for the two populations. For each experiment, the reported results (see Table 6.3) are based on 100 simulations as before. Like the equal prior case, LDA led to the best performance in the case of normal distributions, while error rates of the two other classifiers were also fairly comparable. Once again, in the case of Cauchy distributions, the modified depth classifier outperformed the traditional approaches of linear and quadratic discriminant analysis.

Distribution	μ	π_1	Bayes risk	n	LDA	QDA	HD^*
Normal	1	0.6	23.11	100	23.52(0.14)	23.59(0.14)	25.07(0.17)
				200	23.24(0.12)	23.30(0.13)	24.39(0.14)
		0.7	20.42	100	20.72(0.13)	20.89(0.14)	22.87(0.17)
				200	20.56(0.11)	20.56(0.11)	22.04(0.14)
	2	0.6	7.65	100	7.89(0.08)	7.95(0.05)	8.73(0.12)
				200	7.65(0.09)	7.71(0.09)	8.17(0.09)
		0.7	7.01	100	7.15(0.09)	7.21(0.08)	8,29(0.13)
				200	7.11(0.08)	7.14(0.09)	7.72(0.10)
Cauchy	1	0.6	28.89	100	40.23(0.04)	45.92(0.81)	33.38(0.30)
				200	40.17(0.03)	49.16(0.89)	32.56(0.21)
		0.7	25.01	100	30.34(0.04)	40.45(1.48)	30.92(0.68)
				200	30.16(0.03)	44.53(1.69)	29.48(0.21)
	2	0.6	18.77	100	39.53(0.17)	45.86(0.92)	22.19(0.24)
				200	40.16(0.04)	48.25(0.94)	20.90(0.18)
		0.7	16.70	100	30.51(0.06)	40.97(1.55)	20.82(0.24)
				200	30.26(0.03)	44.79(1.75)	19.66(0.19)

Table 6.3 : Misclassification rates (in %) on elliptic distributions when $\pi_1 \neq \pi_2$ (dimension = 2)

6.6.2 Results on "biomedical data"

We also use the "biomedical data set" (described in Section 4.4 of Chapter 4) to compare the performance of different classification methodologies. As before, we removed 15 out of the 209 observations, which have missing values and applied the classification methods on the remaining 194 cases. Training and test sets were formed by random partition, and this partition was carried out 250 times to generate 250 different training and test samples. In each case 100 observations from the first group and 50 from the second were chosen randomly to form the training sample, while the rest of the observations were used as the test set. Average misclassification rates for different classifiers over those 250 samples are reported here along with their corresponding standard errors. For our experiment, we took the sample proportions of the two classes as their prior probabilities. In this data set, QDA led to the best error rate of 12.26% (with S.E.= 0.21%). The usual k-nearest neighbor classifier based on Euclidean distance and cross-validated estimate of k had an error rate of 13.22% with a standard error of 0.27%. The performance of the HD^* was better than that of LDA. These two classifiers could achieve error rates of 13.98% and 15.75%, respectively, with corresponding standard errors of 0.31% and 0.32% in the respective cases.

6.7 Remarks and discussions

Among the different depth based classifiers discussed in this chapter, SPD seems to be the advantageous one. Not only it requires less computation but also gets less affected by data sparsity in higher dimensions. In most of the examples that we have considered here, SPD led to better performance than the other maximum depth classifiers.

Traditional LDA and QDA are mainly motivated by the normality of the data distribution and highly sensitive to outliers and extreme values. But the depth based classifiers discussed here are robust in nature and works well even when the observations come from distributions having heavy tails.

Maximum depth classifiers differ critically from the depth based semiparametric classifiers discussed in Chapter 5. Results on the performance of depth based semiparametric classifiers have been reported in the preceding chapter for simulated as well as all of the benchmark data sets discussed here except the salmon data. For salmon data, we observed the error rates of 7.17% (with S.E.= 0.34%) and 5.88% (with S.E.=0.30%), respectively, when the half-space depth and the regression depth were used to construct the linear separating surfaces. Depth based quadratic classifiers had a slightly higher error rates in this data set.

Compared to the performance of the other well known nonparametric methods like

nearest neighbors and classification trees, maximum depth classifiers, specially the SPD classifier, led to fairly satisfactory performance in all the data sets that we have analyzed here. From the numerical results it seems that SPD classifier has the potential to be used as a computationally efficient robust alternative for the traditional methods of discriminant analysis.

6.8 Proofs and mathematical details

Proposition 6.1 : If the density $f(\mathbf{x})$ of a spherically symmetric distribution (in dimension ≥ 2) is strictly decreasing in distance from the center of symmetry, so is the spatial depth.

Proof of Proposition 6.1 : Without loss of generality, we can take the origin as the point of symmetry. As f is spherically symmetric, it is easy to see that the points at the same distance from the center have the same spatial depth since it is invariant under orthogonal transformation. Now, choose two points $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\|\mathbf{x}_1\| < \|\mathbf{x}_2\|$ (i.e. $f(\mathbf{x}_1) > f(\mathbf{x}_2)$). Because of spherical symmetry, without loss of generality we can take the observation on the same co-ordinate axis. Let $\mathbf{x}_1 = (t_1, 0, \ldots, 0)$ and $\mathbf{x}_2 = (t_2, 0, \ldots, 0)$ where $|t_1| < |t_2|$. Next, notice that for any observation $\mathbf{x}^{(1)} = (x_1, x_2, \ldots, x_d)$, it is possible to find three other points $\mathbf{x}^{(2)} = (x_1, -x_2, -x_3, \ldots, -x_d)$, $\mathbf{x}^{(3)} = (-x_1, x_2, x_3, \ldots, x_d)$ and $\mathbf{x}^{(4)} = (-x_1, -x_2, -x_3, \ldots, -x_d)$ such that $f(\mathbf{x}^{(1)}) = f(\mathbf{x}^{(2)}) = f(\mathbf{x}^{(3)}) = f(\mathbf{x}^{(4)})$, and both $\sum_{i=1}^{4} \frac{\mathbf{x}^{(i)} - \mathbf{x}_1}{\|\mathbf{x}^{(i)} - \mathbf{x}_1\|} f(\mathbf{x}^{(i)})$ and $\sum_{i=1}^{4} \frac{\mathbf{x}^{(i)} - \mathbf{x}_2}{\|\mathbf{x}^{(i)} - \mathbf{x}_2\|} f(\mathbf{x}^{(i)})$ are vectors along that co-ordinate axis directed towards the origin with the second one having a larger magnitude. Now, taking integration over all such $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}$, we obtain $\left\| E_{\mathbf{X}} \left\{ \frac{\mathbf{x}_1 - \mathbf{X}}{\|\mathbf{x}_1 - \mathbf{X}\|} \right\} \right\| < \left\| E_{\mathbf{X}} \left\{ \frac{\mathbf{x}_2 - \mathbf{X}}{\|\mathbf{x}_2 - \mathbf{X}\|} \right\} \right\|$. Hence the proof.

In order to prove Theorems 6.1-6.3, first note that

$$|\Delta_{\mathbf{n}} - \Delta| \leq \sum_{j=1}^{J} \pi_j \int \left| \prod_{\substack{i=1 \\ i \neq j}}^{J} I\{D_{n_j}(j, \mathbf{x}) > D_{n_i}(i, \mathbf{x})\} - \prod_{\substack{i=1 \\ i \neq j}}^{J} I\{D(j, \mathbf{x}) > D(i, \mathbf{x})\} \right| f_j(\mathbf{x}) d\mathbf{x},$$

where $D_{n_j}(j, \mathbf{x})$ and $D(j, \mathbf{x})$ are the empirical depth and population depth of $\mathbf{x}$ in the j^{th} population ($j = 1, 2, \ldots, J$), π_j's are the prior probabilities and f_j's are the density functions of the respective classes. Therefore, if one can show the pointwise convergence of empirical depth functions to population depth, the result will follow as an immediate consequence of Dominated Convergence Theorem, when the population depth based classifiers are the optimal Bayes classifier. Note that, for HD, SD, MJD and PD, the population depth based classifiers are the Bayes classifier, when the populations are elliptic differing only in their location parameters, and for SVD^{δ}, if one has the additional condition $\delta \geq 1$ the same

assertion holds (see e.g., Zuo and Serfling, 2000a). The population version of SPD leads to the Bayes classifier under the condition of location shift and spherical symmetry.

Proof of Theorem 6.1 : Here, instead of pointwise convergence, in fact, the results on uniform convergence of the empirical versions HD, SD, MJD and PD are well known in the literature (see e.g., Nolan, 1992; Donoho and Gasko, 1992; Liu, 1990; Liu and Singh, 1993; Zuo and Serfling, 2000b).

Proof of Theorem 6.2 : Here also, we have results on uniform convergence of the empirical depth function. Uniform convergence of the empirical version of spatial depth to its population analogue follows from the work of Kolchinskii (1997) and Serfling (2002).

Proof of Theorem 6.3 : Since the populations satisfy the location shift model, it is not necessary to have $|\Sigma|^{1/2}$ in the denominator of the expression of SVD^{δ}, and it can be ignored. Now, under the assumed condition, it follows from the result on U-statistic that for any given $\mathbf{x}$, empirical version of SVD^{δ} converges almost surely to population counter part.

Proof of Theorem 6.4 : Note that under the given conditions the population depth based classifiers turn out to be the optimal Bayes classifier, and therefore Δ can be expressed as

$$\Delta = J^{-1}\sum_{j=1}^{J} P\{\arg\max_{k} D(k,\mathbf{x}) \neq j \text{ when actually } \mathbf{x} \in j\text{-th population}\}$$

(i) (The case of HD) : From Hoeffding's (1963) lemma for i.i.d. random variables, for any fixed $\mathbf{x}$ and l and for every $\epsilon > 0$, we have

$$P\left\{\left| n_j^{-1}\sum_{i=1}^{n_j} I\{l'(\mathbf{x}_{ji}-\mathbf{x})>0\} - P\{l'(\mathbf{X}_j-\mathbf{x})>0\}\right| > \epsilon\right\} < 2e^{-2n_j\epsilon^2} \text{ for } j=1,2,\ldots,J.$$

Now, for some fixed $\mathbf{x}$, the sets of hyperplanes $\{l'(\mathbf{X}-\mathbf{x})=0\}$ in R^d has VC dimension d (see e.g., Pollard, 1984). So, the sets of the form $\{\mathbf{X} : l'(\mathbf{X}-\mathbf{x})>0\}$ have polynomial discrimination with d being the degree of the polynomial. Therefore, using the results on probability inequalities on such sets (see e.g., Pollard, 1984), for $j=1,2,\ldots,J$, and every $\epsilon>0$, we get

$$P\left\{\sup_{l}\left| n_j^{-1}\sum_{i=1}^{n_j} I\{l'(\mathbf{x}_{ji}-\mathbf{x})>0\} - P\{l'(\mathbf{X}_j-\mathbf{x})>0\}\right| > \epsilon\right\} < 2\, n_j^{d} e^{-2n_j\epsilon^2}.$$

Again, $\left|\sup_{l} n_j^{-1}\sum_{i=1}^{n_j} I\{l'(\mathbf{x}_{ji}-\mathbf{x})>0\} - \sup_{l} P\{l'(\mathbf{X}_j-\mathbf{x})>0\}\right| > \epsilon$

$$\Rightarrow \sup_{l}\left| n_j^{-1}\sum_{i=1}^{n_j} I\{l'(\mathbf{x}_{ji}-\mathbf{x})>0\} - P\{l'(\mathbf{X}_j-\mathbf{x})>0\}\right| > \epsilon.$$

Therefore, $P\left\{\left|D_{n_j}(j,\mathbf{x}) - D(j,\mathbf{x})\right| > \epsilon\right\} < 2\, n_j^d e^{-2n_j\epsilon^2}$. Now, assume that $D^{01}(\mathbf{x}) = \min_{\{j\,:\,j\neq 1\}}\{D(1,\mathbf{x}) - D(j,\mathbf{x})\} > 0$ and choose $\epsilon = D^{01}(\mathbf{x})/2$.

$$\begin{aligned} P\{D_{\mathbf{n}}^{01}(\mathbf{x}) > 0\} &\geq P\{|D_{n_j}(j,\mathbf{x}) - D(j,\mathbf{x})| < D^{01}(\mathbf{x})/2 \text{ for every } j = 1,2,\ldots,J.\} \\ &\geq \prod_{j=1}^{J} max\{0, 1 - 2n_j^d e^{-n_j[D^{01}(\mathbf{x})]^2/2}\} = \beta_{\mathbf{n}}^*\{D^{01}(\mathbf{x})\}, \text{say.} \end{aligned}$$

Clearly, $\beta_{\mathbf{n}}^*\{D^{01}(\mathbf{x})\} > 0$ and $P\{D_{\mathbf{n}}^{01}(\mathbf{x}) < 0\} \leq 1 - \beta_{\mathbf{n}}^*\{D^{01}(\mathbf{x})\}$

$$\begin{aligned} \Rightarrow J(\Delta_{\mathbf{n}} - \Delta) &= \sum_{j=1}^{J} \int_{D^{0j}(\mathbf{x})>0} P\{D_n^{0j}(\mathbf{x}) < 0\} f_j(\mathbf{x}) d\mathbf{x} \\ &< \sum_{j=1}^{J} \int_{D^{0j}(\mathbf{x})>0} [1 - \beta_{\mathbf{n}}^*\{D^{0j}(\mathbf{x})\}] f_j(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

(*ii*) (The case of SD) : The sample version of simplicial depth is a U-statistic with a bounded kernel function. Therefore, using Hoeffding's (1963) inequality we have,

$$P\left\{\left|D_{n_j}(j,\mathbf{x}) - D(j,\mathbf{x})\right| > \epsilon\right\} < 2\, e^{-2\lfloor n_j/(d+1)\rfloor\epsilon^2} \text{ for every } \epsilon > 0 \text{ and } j = 1,2,\ldots,J.$$

Now, using similar arguments and similar choice of ϵ as used in the case of HD above, we get

$$J(\Delta_{\mathbf{n}} - \Delta) < \sum_{j=1}^{J} \int_{D^{0j}(\mathbf{x})>0} [1 - \beta_{\mathbf{n}}^{\circ}\{D^{0j}(\mathbf{x})\}] f_j(\mathbf{x}) d\mathbf{x}, \text{ for } \beta_{\mathbf{n}}^{\circ}(t) = \prod_{j=1}^{J} max\{0, 1 - 2e^{-\lfloor n_j/d+1\rfloor t^2/2}\}.$$

(*iii*) (The case of SPD) : For the ease for notation, let us define $\mathbf{z}_i = (\mathbf{x} - \mathbf{x}_{ji})/\|\mathbf{x} - \mathbf{x}_{ji}\|$ for $i = 1,2,\ldots,n_j$ and $\mathbf{Z} = (\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|$, where $\mathbf{X} \sim f_j$. Also define $\bar{\mathbf{z}}_{n_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{z}_i$ and $\mu_{\mathbf{Z}} = E(\mathbf{Z})$. Since, $\|\bar{\mathbf{z}}_{n_j}\|$ and $\|\mu_{\mathbf{Z}}\|$ both are positive, we have

$$P\left\{\left|\|\bar{\mathbf{z}}_{n_j}\| - \|\mu_{\mathbf{Z}}\|\right| > \epsilon\right\} < \sum_{k=1}^{d} P\left\{\left|\bar{\mathbf{z}}_{n_j}^2(k) - \mu_{\mathbf{Z}}^2(k)\right| > \epsilon^2/d\right\},$$

where $\bar{\mathbf{z}}_{n_j}(k)$ and $\mu_{\mathbf{Z}}(k)$ are the k^{th} components of $\bar{\mathbf{z}}_{n_j}$ and $\mu_{\mathbf{Z}}$ respectively. Now, for every $k = 1,2,\ldots,d$, we have

$$P\left\{\left|\bar{\mathbf{z}}_{n_j}^2(k) - \mu_{\mathbf{Z}}^2(k)\right| > \epsilon^2/d\right\} < P\left\{\left|\bar{\mathbf{z}}_{n_j}(k) - \mu_{\mathbf{Z}}(k)\right| > \epsilon^2/2d\right\} \text{ (since } |\bar{\mathbf{z}}_{n_j}(k) + \mu_{\mathbf{Z}}(k)| \leq 2)$$

Since $\bar{\mathbf{z}}_{n_j}(k)$ is an average of $i.i.d$ bounded random variables (bounded between -1 and 1), using Hoeffding's lemma we get

$$P\left\{\left|\bar{\mathbf{z}}_{n_j}(k) - \mu_{\mathbf{Z}}(k)\right| > \epsilon^2/2d\right\} < 2e^{-n_j\epsilon^4/8d^2}$$

$$\Rightarrow P\{|D_{n_j}(j,\mathbf{x}) - D(j,\mathbf{x})| > \epsilon\} = P\left\{\left|\|\bar{\mathbf{z}}_{n_j}\| - \|\mu_{\mathbf{Z}}\|\right| > \epsilon\right\} = 2d\ e^{-n_j\epsilon^4/8d^2}.$$

Now, using similar arguments and similar choice of ϵ as used in the other two cases, we obtain

$$J(\Delta_{\mathbf{n}} - \Delta) < \sum_{j=1}^{J} \int_{D^{0j}(\mathbf{x})>0} [1 - \beta_{\mathbf{n}}^{+}\{D^{0j}(\mathbf{x})\}] f_j(\mathbf{x})d\mathbf{x}, \quad \text{for } \beta_{\mathbf{n}}^{+}(t) = \prod_{j=1}^{J} max\{0, 1 - 2d\ e^{-n_j t^4/8d^2}\}.$$

Proof of Theorem 6.5 : Let μ_j and Σ_j be the location parameter and scatter matrix of the j^{th} population which has a density function g_j. Define $R_j = \left\{(\mathbf{X}_j - \mu_j)'\Sigma_j^{-1}(\mathbf{X}_j - \mu_j)\right\}^{1/2}$ where $\mathbf{X}_j \sim g_j$. When, g_j is elliptically symmetric, the distributions of R_j is given by (see e.g., Fang, Kotz and Ng, 1989)

$$\varrho_j(r_j) = \frac{\pi^{d/2}}{\Gamma d/2}|\Sigma_j|^{1/2} r_j^{d-1} g_j(\mathbf{x}), \quad 0 < r_j < \infty,$$

where $r_j = \{(\mathbf{x} - \mu_j)'\Sigma_j^{-1}(\mathbf{x} - \mu_j)\}^{1/2}$, the Mahalanobis distance of $\mathbf{x}$ from μ_j. Clearly, $\pi_j g_j(\mathbf{x}) > \pi_i g_i(\mathbf{x}) \Leftrightarrow \pi_j|\Sigma_j|^{-1/2}\ \varrho_j(r_j)/r_j^{d-1} > \pi_i|\Sigma_i|^{-1/2}\ \varrho_i(r_i)/r_i^{d-1}$, and one should also notice that in the case of elliptic populations the Mahalanobis distance r_j is a function of population depth $D(j,\mathbf{x})$. Let us define $r_j = \gamma_j\{D(j,\mathbf{x})\}$. Now, it is easy to see that the optimal Bayes classifier can be given by

$$\mathrm{d}_B(\mathbf{x}) = \arg\max_j \pi_j\theta_j\{D(j,\mathbf{x})\}, \quad \text{where } \theta_j(t) = |\Sigma_j|^{-1/2}\ \varrho_j\{\gamma_j(t)\}/\{\gamma_j(t)\}^{d-1}.$$

Lemma 6.1 : Suppose that $\mathbf{X}$ follows an elliptic distribution G, and it is distributed as $\mathbf{X} = \mu + \Sigma^{1/2}\mathbf{U}$, where μ and Σ are some location and scale parameters, and $\mathbf{U}$ is a spherically distributed random variable. Then, for any given $\mathbf{x}$ and any given direction α ($\|\alpha\| = 1$), we have

$$[\xi_{1-\delta,\alpha} - \xi_{\delta,\alpha}]/2 = \{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2}\sqrt{\alpha'\Sigma\alpha},$$

where $\delta = HD(G,\mathbf{x})$, the half-space depth of $\mathbf{x}$ with respect to G, and $\xi_{p,\alpha}$ is the p-th quantile of $\alpha'\mathbf{X}$.

Proof of Lemma 6.1 : Half-space depth of $\mathbf{x}$ w.r.t. G can be expressed as

$$HD(G,\mathbf{x}) = 1 - \sup_{\alpha} P\{\alpha'(\mathbf{X} - \mathbf{x}) < 0\} = 1 - \sup_{\alpha} P\left\{\frac{\alpha'(\mathbf{X} - \mu)}{\sqrt{\alpha'\Sigma\alpha}} < \frac{\alpha'(\mathbf{x} - \mu)}{\sqrt{\alpha'\Sigma\alpha}}\right\}$$

It is easy to check that $\frac{\alpha'(\mathbf{X}-\mu)}{\sqrt{\alpha'\Sigma\alpha}}$ is distributed as $l'\mathbf{U}$ with $\|l\| = 1$. Therefore,

$$\begin{aligned}\delta = HD(G,\mathbf{x}) &= 1 - \sup_{\alpha} F\left[\frac{\alpha'(\mathbf{x} - \mu)}{\sqrt{\alpha'\Sigma\alpha}}\right] \\ &= 1 - F\left[\sup_{\alpha}\frac{\alpha'(\mathbf{x} - \mu)}{\sqrt{\alpha'\Sigma\alpha}}\right] = 1 - F\left[\{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2}\right],\end{aligned}$$

where F is the distribution function of $l'\mathbf{U}$ for every l with $\|l\| = 1$. Now, from the definition of $\xi_{p,\alpha}$, it is quite clear that

$$P\{\alpha'\mathbf{X} < \xi_{p,\alpha}\} = F\left(\frac{\xi_{p,\alpha} - \alpha'\mu}{\sqrt{\alpha'\Sigma\alpha}}\right) = p.$$

Taking $p = 1 - \delta$ and $p = \delta$, we get

$$F\left(\frac{\xi_{1-\delta,\alpha} - \alpha'\mu}{\sqrt{\alpha'\Sigma\alpha}}\right) = 1 - \delta = F\left[\{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2}\right] \quad \text{and}$$
$$F\left(\frac{\xi_{\delta,\alpha} - \alpha'\mu}{\sqrt{\alpha'\Sigma\alpha}}\right) = \delta = 1 - F\left[\{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2}\right] = F\left[-\{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2}\right]$$

Now, since F is strictly monotonic, we have $\frac{\xi_{1-\delta,\alpha} - \alpha'\mu}{\sqrt{\alpha'\Sigma\alpha}} = -\frac{\xi_{\delta,\alpha} - \alpha'\mu}{\sqrt{\alpha'\Sigma\alpha}} = \{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2} \Rightarrow [\xi_{1-\delta,\alpha} - \xi_{\delta,\alpha}]/2 = \{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2}\sqrt{\alpha'\Sigma\alpha}$.

Lemma 6.2 : Suppose that ζ_p is the unique solution of $F(\mathbf{x}) = p$ $(0 < p < 1)$, and $\zeta_{p,n}$ is its empirical version based on F_n, the empirical distribution function. Also assume that δ and δ_n are the half-space depths of an observation $\mathbf{x}$ with respect to F and F_n, respectively. Then, as $n \to \infty$, $|\zeta_{\delta_n,n} - \zeta_\delta| \xrightarrow{a.s} 0$.

Proof of Lemma 6.2 : Since, ζ_p is a continuous function of p, for every $\epsilon > 0$, there exists an $\eta > 0$ such that $|\delta_n - \delta| < \eta \Rightarrow |\zeta_{\delta_n} - \zeta_\delta| < \epsilon/2$. Therefore, $P\{|\zeta_{\delta_n} - \zeta_\delta| > \epsilon/2\} < P\{|\delta_n - \delta| > \eta\} < 2n^d e^{-2n\eta^2}$. Now, from a theorem in Serfling (1980, pp. 75-76), it follows that for every δ_n $(0 < \delta_n < 1)$, $P\{|\zeta_{\delta_n,n} - \zeta_{\delta_n}| > \epsilon/2\} < 2e^{-2na_n^2}$ where $a_n = min\{F(\zeta_{\delta_n} + \epsilon/2) - \delta_n, \delta_n - F(\zeta_{\delta_n} - \epsilon/2)\}$. Therefore,

$$P\{|\zeta_{\delta_n,n} - \zeta_\delta| > \epsilon\} < 2n^d e^{-2n\eta^2} + 2e^{-2na_n^2}.$$

From the results on convergence of empirical half-space depth (also follows from the proof of part (i) of Theorem 6.4), it is easy to see that $\delta_n \xrightarrow{a.s} \delta$ as $n \to \infty$. So, one can always have an integer n_0 and an interval $I = [\delta - \nu, \delta + \nu]$ $(0 < \delta - \nu < \delta + \nu < 1)$ such that $\delta_n \in I$ for all $n > n_0$. Notice that $\min_n a_n > \inf_{t \in I}[min\{F(\zeta_t + \epsilon/2) - t, t - F(\zeta_t - \epsilon/2)\}] = m$ (say) > 0. Hence, for all $n > n_0$ we have

$$P\{|\zeta_{\delta_n,n} - \zeta_\delta| > \epsilon\} < 2n^d e^{-2n\eta^2} + 2e^{-2nm^2}$$

Now, the result follows from Borel Cantelli Lemma.

Proof of Theorem 6.6 : From Lemma 6.2, it is easy to see that $\xi_{\delta_n,\alpha,n} \xrightarrow{a.s} \xi_{\delta,\alpha}$ and $\xi_{1-\delta_n,\alpha,n} \xrightarrow{a.s} \xi_{1-\delta,\alpha}$, where $\xi_{p,\alpha}$ is the p-th quantile of $\alpha'\mathbf{X}$. Now, from Lemma 6.1, it follows that

$$(\xi_{1-\delta_n,\alpha,n} - \xi_{\delta_n,\alpha,n})/2 \xrightarrow{a.s} (\xi_{1-\delta,\alpha} - \xi_{\delta,\alpha})/2 = \{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\}^{1/2}\sqrt{\alpha'\Sigma\alpha}.$$

Proposition 6.2 : Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are observations from G (as described in Lemma 6.1), and $\xi_{p,\alpha}$ and $\xi_{p,\alpha,n}$ have the same meaning as in Theorem 6.6. For some given α ($\|\alpha\| = 1$), define $v_i = \{(\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu)\}^{1/2} \sqrt{\alpha' \Sigma \alpha}$ and $\hat{v}_i = (\xi_{1-\delta_{i_n},\alpha,n} - \xi_{\delta_{i_n},\alpha,n})/2$ for $i = 1, 2, \ldots, n$, where δ_{i_n} is the empirical half-space depth of $\mathbf{x}_i$. Similarly, define v_0 and $\hat{v}_0$ for a new observation $\mathbf{x}_0$. Assume that the v_i's have density function Υ and define its kernel density estimate $\hat{\Upsilon}^*_{h_n}(v) = \frac{1}{nh_n} \sum_{i=1}^{n} K\{(v - \hat{v}_i)/h_n\}$ for some kernel function K and bandwidth $h_n > 0$. Further, assume that Υ, K and h_n satisfy the following conditions :-

(*i*) Υ has bounded third derivative,

(*ii*) K is symmetric, it has bounded first derivative and $\int |t|^3 K^2(t) dt < \infty$,

(*iii*) $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$.

Then, $\hat{\Upsilon}^*_{h_n}(\hat{v}_0)$ converges to $\Upsilon(v_0)$ in probability as $n \to \infty$.

Proof of Proposition 6.2 : Define $\hat{\Upsilon}_{h_n}(v) = \frac{1}{nh_n} \sum_{i=1}^{n} K\{(v - v_i)/h_n\}$. From the definitions of $\hat{\Upsilon}_{h_n}(.)$ and $\hat{\Upsilon}^*_{h_n}(.)$, it is easy to see that, for every $\epsilon > 0$,

$$P\{|\hat{\Upsilon}_{h_n}(v_0) - \hat{\Upsilon}^*_{h_n}(\hat{v}_0)| > \epsilon\} < nP\{|K\{(v_0 - v_1)/h_n\} - K\{(\hat{v}_0 - \hat{v}_1)/h_n\}| > h_n\epsilon\}.$$

Now, $K\{(v_0 - v_1)/h_n\} - K\{(\hat{v}_0 - \hat{v}_1)/h_n\} = \frac{1}{h_n}\{(v_0 - v_1) - (\hat{v}_0 - \hat{v}_1)\} K'(v/h_n)$, for some v lying between $(v_0 - v_1)$ and $(\hat{v}_0 - \hat{v}_1)$. Therefore, when $K'(\cdot)$ is bounded by M, we have

$$P\{|\hat{\Upsilon}_{h_n}(v_0) - \hat{\Upsilon}^*_{h_n}(\hat{v}_0)| > \epsilon\} < nP\{|(v_0 - v_1) - (\hat{v}_0 - \hat{v}_1)| > h_n M\epsilon\} < 2nP\{|v_0 - \hat{v}_0| > h_n M\epsilon/2\}$$

Now, using Lemma 6.1, it is easy to check that $|v_0 - \hat{v}_0| \leq \frac{1}{2}\{|\xi_{\delta,\alpha} - \xi_{\delta_n,\alpha,n}| + |\xi_{1-\delta,\alpha} - \xi_{1-\delta_n,\alpha,n}|$. Therefore,

$$P\{|v_0 - \hat{v}_0| > h_n M\epsilon/2\} < 2P\{|\xi_{\delta,\alpha} - \xi_{\delta_n,\alpha,n}| > h_n M\epsilon/2\} < 4(n^d e^{-nh_n^2\eta^2/2} + e^{-nh_n^2 m^2/2})$$

for some $\eta > 0$ and $m > 0$ as chosen in Lemma 6.2. This implies that

$$P\{|\hat{\Upsilon}_{h_n}(v_0) - \hat{\Upsilon}^*_{h_n}(v_{0_n})| > \epsilon\} < 8(n^{d+1} e^{-nh_n^2\eta^2/2} + e^{-nh_n^2 m^2/2})$$

and hence $|\hat{\Upsilon}^*_{h_n}(\hat{v}_0) - \hat{\Upsilon}_{h_n}(v)| \xrightarrow{P} 0$ as $n \to \infty$. Now, under the assumed conditions, the expectation and the variance of $\hat{\Upsilon}_{h_n}(v)$ are of the form (see Lemma 2.2)

$$E\{\hat{\Upsilon}_{h_n}(v)\} = \Upsilon(v) + O(h_n^2) \quad \text{and} \quad Var\{\hat{\Upsilon}_{h_n}(v)\} = O(n^{-1}h_n^{-1}),$$

which implies that $|\hat{\Upsilon}_{h_n}(v) - \Upsilon(v)| \xrightarrow{P} 0$ as $n \to \infty$. Therefore, $\hat{\Upsilon}^*_{h_n}(\hat{v}_0)$ converges to $\Upsilon(v)$ in probability.

Chapter 7

Concluding remarks

This thesis deals with some nonparametric and semiparametric classification procedures using kernels, nearest neighbors and data depths. In Chapters 2, 3 and 4 of this thesis, we have proposed some modifications over the existing classification techniques based on kernels and nearest neighbors, while in Chapters 5 and 6 some depth based ideas have been used to construct semiparametric and nonparametric classification rules. Here we point out some relevant problems that arise in course of this research, but have not been fully solved in this thesis. We intend to investigate these problems in future.

In all the examples that we have analyzed in Chapter 2 of this thesis, specially when the prior probabilities of the competing populations are equal, the optimal bandwidth for classification turned out to be larger than that obtained by minimizing the $MISE$ of the kernel density estimate. Our intuition tells that these larger bandwidths try to reduce the variances of the density estimates maintaining the order of their expectations, and thereby reduce the misclassification probabilities. However, further theoretical investigation is needed for proper understanding of this phenomenon.

In Chapter 3 of this thesis, for a fixed pair of bandwidths (h_1, h_2), we have studied the asymptotic properties of the discrimination measures like posterior probabilities and p-values, but no such analysis have been carried out to study the large sample behavior of the entire plot of discrimination measures $\{\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x}) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$ or $\{P_{h_1,h_2}(\mathbf{x}) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$ viewed as a stochastic process indexed by two bandwidth parameters. Some theoretical investigation is needed in this regard. The aggregation procedure adopted for combining the classification results is also somewhat subjective, and one may use many other aggregation procedures as well. It is not known at this moment how to obtain the optimal weighting scheme for a given problem.

In Chapter 4 of this thesis, though we have studied some properties of nearest neighbor based discrimination measures like posterior probabilities and Bayesian measures

of strength, some other interesting aspects of asymptotic behavior of the discrimination measures are yet to be investigated. For instance, it would be useful to study the asymptotics of the discrimination measures viewed as stochastic processes indexed by the neighborhood parameter. In this chapter, following the idea of multi-scale kernel discriminant analysis we have developed another visualization and aggregation procedure for classifiers based on nearest neighbor density estimates. However, unlike what was done in Chapter 3, here we could not find any p-value type easily computable measure for finding the evidence in favor of different populations.

In Chapter 5 of this thesis, to solve the multi-class problems, we had to adopt a pairwise classification procedure followed by majority voting or pairwise coupling. In the trivial situation when the populations are completely separated, one can use suitable optimization algorithms (see e.g., Duda *et. al.*, 2000) for direct estimation of discriminating surfaces. However, for the case of overlapping populations, such a construction is not yet available. We have observed that the error rate of the regression depth classifier always converges to that of the best linear (or quadratic) classifier, but for the convergence of the classifier itself we need a uniqueness condition. We have derived some conditions on the population distributions under which the best linear (or quadratic) classifier is unique, and it is the optimal Bayes classifier. But we believe that there is scope for improving these results. Also note that half space depth and regression depth classifiers use two different criteria for optimization. In the case of linear classification, we have found some conditions for equivalence of these two criteria. One may look for the generalization of these results for nonlinear classification problems. Because of the difficulty in computing exact depth functions, for all high dimensional classification problems we adopted the approximate versions of half space depth and regression depth functions. One can think for some other alternatives in this regard. Moreover, the possible use of other depth functions for construction of separating surfaces is yet to be investigated.

In Chapter 6 of this thesis, due to the difficulty in computing many of the depth functions, we could not carry out data analytic studies for the classifiers based on those depths, especially in high dimensions. If some good approximate but fast algorithms for those depth functions can be developed, it would help in further investigation of depth based classifiers. Also, for the unequal prior case, we could only generalize the classifier based on half space depth. Similar generalization of classifiers based on other depth functions in unequal prior problems remains an open problem at this moment.

We have proposed several classification techniques in this thesis and investigated their statistical properties. At this point, the question that naturally arises is given a specific data set which method should one use for classification. There is no universal objective answer to this question, and the choice of the classifier depends on the data set in

hand as well as various other issues. Every method has its own merits and demerits, and depending on the distributional geometry of the data cloud, one or several of them may appear as the best classifier for that data set. Depth based semiparametic classifiers will be preferred in the cases where the discriminating surfaces can be well approximated by known finite dimensional parametric functions (e.g., linear or quadratic functions). If we have any prior idea about functional form of the separating surface, these methods are more useful than other nonparametric classifiers. In all other cases, one should look for nonparametric methods for classification. Kernel, nearest neighbor and maximum depth classifiers, the three nonparametric methods investigated in this thesis, could produce fairly competitive performance on the benchmark data set analyzed here. In a specific problem, one may use resampling techniques like the bootstrap or cross validation to estimate the misclassification rates for selecting one of these three classifiers. Because of the computational efficiency, one should prefer the spatial depth classifier among different maximum depth classifiers when computing time is a major concern. But, one should also note that its use is restricted to the equal prior case while the classifier based on half-space depth, which is computationally more expensive than spatial depth, can be used for problems with unequal priors as well.

In practice, instead of considering any single classifier, it is always a better idea to run a number of classifiers on a data set and analyze their results before taking the final decision. One can also combine the results from different classifiers using a multiple classifier system (see e.g., Ho, Hull and Srihari, 1994; Breiman, 1996; Friedman, Hastie and Tibshirani, 2000), which generally performs better than a single classification method.

Bibliography

1. Aeberhard, S., Coomans, D. and de Vel, 0. (1994) Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition*, 27, 1065-1077.

2. Albert, A. and Anderson, J. A. (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1-10.

3. Alpaydin, E. (1997) Voting over multiple condensed nearest neighbor *Artificial Intelligence Review*, 11, 115-132.

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

5. Bai, Z.-D. and He, X. (1999) Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *The Annals of Statistics*, 27, 1616-1637.

6. Bensmail, H., and Bozdogan, H. (2002). Model-based kernel discriminant analysis with optimal scaling. In press for *The Institute of Statistical Mathematics in Japan*, Springer-Verlag, Tokyo.

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

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

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

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

11. Breiman, L. (1998) Arcing classifiers (with discussion) *The Annals of Statistics*, 26, 801-849.

12. Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2, 121-167.

13. Campbell, N. A. and Mahon, R. J. (1974) A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Australian Journal of Zoology*, 22, 417-425.

14. Chaudhuri, P. and Sengupta, D. (1993) Sign tests in multi-dimension : inference based on the geometry of the data cloud. *Journal of the American Statistical Association*, 88, 1363-1370.

15. Chaudhuri, P. (1996) On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, **91**, 862-872.

16. Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807-823.

17. Chaudhuri, P. and Marron, J. S. (2000) Scale space view of curve estimation. *The Annals of Statistics*, **28**, 408-428.

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

19. Christmann, A. and Rousseeuw, P. (2001) Measuring overlap in binary regression. *Computational Statistics and Data Analysis*, **37**, 65-75.

20. Christmann, A. (2002) Classification based on support vector machine and on regression depth. *Statistics in Industry and Technology : Statistical Data Analysis* (Edited by : Y. Dodge), Birkhauser, Berlin, pp. 341-352.

21. Christmann, A., Fischer, P. and Joachims, T. (2002) Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, **17**, 273-287.

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

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

24. Cover, T. M. and Hart, P. E. (1968) Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, **13**, 21-27.

25. 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.

26. Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge.

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

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

29. Devroye, L. (1981) On the almost everywhere convergence of nonparametric regression function estimates. *The Annals of Statistics*, **9**, 1310-1319.

30. Devroye, L., Gyorfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

31. Donoho, D. (1982) Breakdown properties of multivariate location estimators. *Ph.D. qualifying paper, Department of Statistics, Harvard University.*

32. Donoho, D. and Gasko, M. (1992) Breakdown properties of location estimates based on half-space depth and projected outlyingness. *The Annals of Statistics*, 20, 1803-1827.

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

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

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

36. Fang, K-T., Kotz, S. and Ng, K. W. (1989) *Symmetric Multivariate and Related Distributions.* Chapman and Hall, London.

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

38. Fix, E. and Hodges, J. L., Jr.,(1951) Discriminatory analysis, nonparametric discrimination, consistency properties. *Randolph Field, Texas, Project 21-49-004, Report No. 4.*

39. Friedman, J. H. (1989) Regularized discriminant analysis. *Journal of the American Statistical Association*, 84, 165-175.

40. Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, 19, 1–141.

41. Friedman, J. H. (1994) Flexible metric nearest neighbor classification. *Technical Report, Department of Statistics, Stanford University.*

42. Friedman, J. H. (1996) Another approach to ploychotomous classification. *Technical Report, Department of Statistics, Stanford University.*

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

44. Friedman, J. H., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression : a statistical view of boosting (with discussion). *The Annals of Statistics*, 28, 337-374.

45. Fruend, Y. (1995) Boosting a weak learning algorithm by majority. *Information and Computation*, 121, 256-285.

46. Fruend, Y. and Schapire, R. E. (1997) A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119-139.

47. Fukunaga, K. and Hostetler, L. D. (1973) Optimization of k-nearest neighbor density estimates. *IEEE Transactions on Information Theory*, 19, 320-326.

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

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

50. Ghosh, A. K. and Chaudhuri, P. (2003a) On data depth and distribution free discriminant analysis using separating surfaces. To appear in *Bernoulli.*

51. Ghosh, A. K. and Chaudhuri, P. (2003b) On maximum depth classifiers. *Submitted for publication.*

52. Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2003a) Multi-scale kernel discriminant analysis. *Proceedings of Fifth International Conference on Advances in Pattern Recognition ICAPR-03* (Edited by : D. P. Mukherjee and S. Pal), Allied Publishers, Kolkata, pp. 89-93.

53. Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2003b) Classification using kernel density estimates : multi-scale analysis and visualization. Tentatively accepted for publication in *Technometrics.*

54. Ghosh, A. K., Chaudhuri, P. and Murthy, C. A: (2003) On visualization and aggregation of nearest neighbor classifiers. *Submitted for publication.*

55. Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

56. Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2002) Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics*, 11, 1-22.

57. 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.

58. Hall, P. (1983) Large sample optimality of least squares cross validations in density estimation. *The Annals of Statistics*, 11, 1156-1174.

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

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

61. 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.

62. Hand, D. J. (1981) *Discrimination and Classification*, Wiley, New york.

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

64. Hart, P. E. (1968) The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14, 515-516.

65. Hastie, T., Tibshirani, R. and Buja, A. (1994) Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, 1255-1270.

66. Hastie, T. and Tibshirani, R. (1996) Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 607-616.

67. Hastie, T. and Tibshirani, R. (1998) Classification by pairwise coupling. *The Annals of Statistics*, 26, 451-471.

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

69. He, X. and Wang, G. (1997) Convergence of depth contours for multivariate data sets. *The Annals of Statistics*, 25, 495-504.

70. Hills, M. (1966) Allocation rules and their error rates. *Journal of the Royal Statistical Society, Series B*, 28, 1-31.

71. Ho, T. K., Hull, J. J. and Srihari, S. N. (1994) Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 66-75.

72. Hodges, J. (1955) A bivariate sign test. *The Annals of Mathematical Statistics*, 26, 523-527.

73. Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58, 13-30.

74. Holmes, C. C. and Adams, N. M. (2002) A probabilistic nearest neighbor method for statistical pattern recognition. *Journal of the Royal Statistical Society, Series B*, 64, 295-306.

75. Holmes, C. C. and Adams, N. M. (2003) Likelihood inference in nearest-neighbor classification methods. *Biometrika*, 90, 99-112.

76. Huber, P. J. (1985) Projection pursuit (with discussion). *The Annals of Statistics*, 13, 435-475.

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

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

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

80. Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b) A brief summary of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91, 401-407.

81. Jornsten, R., Vardi, Y. and Zhang, C. H. (2002) A robust clustering method and visualization tool based on data depth. *Statistical Data Analysis* (Y. Dodge, ed.). Birkhauser, Basel. 353-366.

82. Jornsten, R. (2004) Clustering and classification based on L_1 data depth. To appear in *Journal of Multivariate Analysis.*

83. Kim, H. and Loh, W.-Y. (2001) Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96, 589-604.

84. Kim, H. and Loh, W.-Y. (2003) Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics*, 12, 512-530.

85. Koltchinskii, V. I. (1997) M-estimation, convexity and quantiles. *The Annals of Statistics*, 25, 435-477.

86. Kooperberg, C., Bose, S. and Stone, C. J. (1997) Polychotomous regression. *Journal of the American Statistical Association*, 92, 117-127.

87. Koshevoy, G. and Mosler, K. (1997) Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25, 1998-2017.

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

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

90. Lim, T. S., Loh, W. Y. and Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40, 203-228.

91. Lippmann, R. P. (1987) An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4, 4-22.

92. Liu, R. (1990) On notion of data depth based on random simplicies. *The Annals of Statistics*, 18, 405-414.

93. Liu, R. and Singh, K. (1993) A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88, 252-260.

94. Liu, R., Parelius, J. and Singh, K. (1999) Multivariate analysis of the data-depth : descriptive statistics and inference. *The Annals of Statistics*, 27, 783-858.

95. Loftsgaarden, D. O. and Quesenberry, C. P. (1965) A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36, 1049-1051.

96. Loh, W.-Y. and Vanichsetakul, N. (1988) Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association*, 83, 715-728.

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

98. Mahalanobis, P. C. (1936) On the generalized distance in statistics. *Proceedings of the National Academy of Sciences, India*, 12, 49-55.

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

100. McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, Chapman and Hall, London.

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

102. Mitra, P., Murthy C. A. and S. K. Pal (2002) Density based multiscale data condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 734-747.

103. Mizera, I. (2002) On depth and deep points : a calculus. *The Annals of Statistics*, 30, 1681-1736.

104. Mizera, I. and Volauf, M. (2002) Continuity of half-space depth contours and maximum depth estimators : diagnostics of depth related methods. *Journal of Multivariate Analysis*, 83, 365-388.

105. Mosler, K. (2002) *Multivariate Dispersions, Central Regions and Depth.* Springer-Verlag, New York.

106. Mosteller, F. and Wallace, D. L. (1963) Inference in an authorship problem. *Journal of the American Statistical Association*, **58**, 275-309.

107. Muller, H. G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics*, **12**, 76-774.

108. Nolan, D. (1992) Asymptotics for multivariate trimming. *Stochastic Processes and Applications*, **42**, 157-169.

109. Oja, H. (1983) Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, **1**, 327-332.

110. Opitz, D. and Maclin, R. (1999) Popular ensemble methods : an empirical study. *Journal of Artificial Intelligence Research*, **11**, 169-198.

111. Pal, S. K., Bandopadhyay, S. and Murthy, C. A. (1998) Genetic algorithms for generation of class boundaries. *IEEE Transactions on Systems, Man and Cybernetics*, **28**, 816-828.

112. Peterson, G. E. and Barney, H. L. (1952) Control methods used in a study of vowels. *The Journal of the Acoustical Society of America*, **24**, 175-185.

113. Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer-Verlag, New York.

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

115. 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.

116. Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion.) *Journal of the Royal Statistical Society, Series B*, **56**, 409-456.

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

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

119. Rousseeuw, P.J. and Ruts, I. (1996) Algorithm AS 307: bivariate location depth. *Applied Statistics*, **45**, 516-526.

120. Rousseeuw, P.J. and Struyf, A. (1998) Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, **8**, 193-203.

121. Rousseeuw, P.J. and Hubert, M. (1999) Regression depth (with discussion). *Journal of the American Statistical Association*, **94**, 388-402.

122. Santner, T. J. and Duffy, D. E. (1986) A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **73**, 755-758.

123. Schapire, R. E., Fruend, Y., Bartlett, P. and Lee, W. (1998) Boosting the margin : a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, **26**, 1651-1686.

124. Scholkopf, S., Burges, C. J. C. and Smola, A. J. (1999) *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge.

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

126. Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

127. Serfling, R. (1992) Nonparametric confidence intervals for generalized quantile parameters in multi-sample contexts. *Nonparametric Statistics and Related Topics* (Edited by : A.K.Md.E. Saleh), Elsevier Science Publishers B.V., pp. 121–139.

128. Serfling, R. (2002) A depth function and a scale curve based on spatial quantiles. *Statistics and Data Analysis based on L_1-Norm and Related Methods* (Edited by : Y. Dodge), Birkhaeuser, Boston, pp. 25-38.

129. Shalak, D. B. (1996) Prototype selections for composite nearest neighbor classifiers. *Ph.D. Thesis, Department of Computer Science, University of Massachusetts.*

130. Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683-690.

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

132. Singh, K. (1991) A notion of majority depth. *Technical Report, Department of Statistics, Rutgers University.*

133. Stahel, W. A. (1981) Breakdown of covariance estimators. *Research Report 31, Fachgruppefur Statistik, ETH, Zurich.*

134. Stone, C. J. (1977) Consistent nonparametric regression (with discussion). *The Annals of Statistics*, **5**, 595-645.

135. Stone, C. J. (1984) An asymptotically optimal window selection rule in kernel density estimates. *The Annals of Statistics*, **12**, 1285-1297.

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

137. Tukey, J. (1975) Mathematics and the picturing of data. *Proceedings of the 1975 International Congress of Mathematics, Vancouver*, pp. 523-531.

138. Vapnik, V. N. and Chervonenkis, A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**, 264-280.

139. Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

140. Vapnik. V. N. (1998) *Statistical Learning Theory.* Wiley, New York.

141. Vardi, Y. and Zhang, C. H. (2000) The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences, U.S.A*, **97**, 1423-1426.

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

143. Zhu, M. and Hastie, T. (2003) Feature extraction for nonparametric discriminant analysis. *Journal of Computational and Graphical Statistics*, **12**, 101-120.

144. Zuo, Y. and Serfling, R. (2000a) General notions of statistical depth function. *The Annals of Statistics*, **28**, 461-482.

145. Zuo, Y. and Serfling, R. (2000b) Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics*, **28**, 483-499.