On Data Depth and Distribution Free Discriminant Analysis Using Separating Surfaces

Anil Kumar Ghosh and Probal Chaudhuri

Theoretical Statistics and Mathematics Unit,

Indian Statistical Institute,

203, B. T. Road, Calcutta-700108, India.

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

Abstract

A very well-known traditional approach in discriminant analysis is to use some linear (or nonlinear) combination of measurement variables which can enhance the class separability. For instance, a linear (or a quadratic) classifier finds out the linear projection (or the quadratic function) of the measurement variables that will maximize 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 multivariate normal distribution, uses the first and the second order moments of the training sample to build such classifiers. These estimates, however, are highly sensitive to outliers, and they are not reliable for heavy tailed distributions. This article investigates two distribution free methods for linear classification, which are based on the notions of statistical depth functions. One of these classifiers is closely related to Tukey's half-space depth (Tukey, 1975) while the other one is based on the concept of regression depth (Rousseeuw and Hubert, 1999). Both these methods can be generalized for constructing nonlinear surfaces to discriminate among competing classes. These depth based methods assume some finite dimensional parametric form of the discriminating surface and use 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.

Keywords : Bayes risk, elliptic symmetry, generalized U-statistic, half-space depth, linear discriminant analysis, location shift models, misclassification rates, optimal Bayes classifier, quadratic discriminant analysis, regression depth, robustness, Vapnik-Chervonenkis dimension.

1 Introduction

Discriminant analysis aims at finding out an appropriate function $f(\mathbf{x})$ of the measurement vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ that contains the maximum information about class separability. In a two-class problem, this function f can be used to construct the separating surface $S = \{\mathbf{x} : f(\mathbf{x}) = 0\}$ between the two classes. For instance, in linear discriminant analysis, one tries to determine a separating hyperplane $S = \{\mathbf{x} : \boldsymbol{\alpha}^{'}\mathbf{x} + \beta = 0\}$ based on the training sample observations. Several methods for choosing the projection vector $\boldsymbol{\alpha}$ and the constant β from the training sample are available in the literature (see e.g., Fukunaga, 1990; McLachlan, 1992; Duda, Hart and Stork, 2000; Hastie, Tibshirani and Friedman, 2001). Similarly, in quadratic discriminant analysis, one uses a quadratic separating surface $S = \{\mathbf{x} : \boldsymbol{\alpha}^{'}\mathbf{x} + \mathbf{x}^{'}\Gamma\mathbf{x} + \beta = 0\}$, where Γ is a symmetric matrix to be chosen from the training sample in addition to $\boldsymbol{\alpha}$ 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 $\boldsymbol{\alpha}$, β and Γ were constructed using the mean vectors and the dispersion matrices of the training samples. Under the assumption of multivariate normal distribution for the data, such linear or quadratic classifiers turn out to be the optimal Bayes classifiers. However, since such methods require the estimation of $\boldsymbol{\alpha}$, β and Γ 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.

In this paper, we will study some linear and nonlinear classification methods that are based on the notions of half-space depth (Tukey, 1975) and regression depth (Rousseeuw and Hubert, 1999). Over the last decade, various notions of data depth have emerged as powerful exploratory and inferential tools for nonparametric multivariate analysis (see e.g., Liu, 1990; Liu, Parelius and Singh, 1999; Vardi and Zhang, 2000; Zuo and Serfling, 2000a; Serfling, 2002; Mosler, 2002). Recently, 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, 1998; Hastie *et. al.*, 2001). Since the discriminant analysis tools investigated in this article are based on half-space 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.

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.

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 $\mathbf{x}$.

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

Sample 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 sample 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, 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 $\mathbf{x}_{1i} - \mathbf{x}_{2j}$ ($\mathbf{x}_{1i}$ and $\mathbf{x}_{2j}$ 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 $\widehat{\boldsymbol{\alpha}}_H$, can be used to construct a linear classifier of the form $\widehat{\boldsymbol{\alpha}}_H^{'}\mathbf{x} + \beta = 0$ 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 $\widehat{\boldsymbol{\alpha}}_H$, $\hat{\beta}_H$ can be obtained by minimizing w.r.t. β the average training set misclassification error $\Delta_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_H, \beta)$ given by the expression

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

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

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 $\boldsymbol{\alpha}_+ = (\boldsymbol{\alpha}, \beta) \in R^{d+1}$ of co-efficients in a linear regression framework. Given a data cloud $\zeta_n = [\{\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id}), y_i\};\ i = 1, 2, \ldots, n]$, $\boldsymbol{\alpha}_+$ is called a 'nonfit' to ζ_n if and only if there exists an affine hyperplane V in the $\mathbf{x}$-space such that no $\mathbf{x}_i$ belongs to V, and the residuals $r_i(\boldsymbol{\alpha}_+) = y_i - \boldsymbol{\alpha}'_+(\mathbf{x}_i, 1)$ are all positive in one open half-space (i.e., one side of V) in the $\mathbf{x}$-space and all negative in the complementary open half-space (i.e., the other side of V). Regression depth of a 'fit' $\boldsymbol{\alpha}_+$ is defined as the minimum number of observations that need to be removed to make it a 'nonfit'.

Recently, Christmann and Rousseeuw (2001) and Christmann *et. al.* (2002) used this notion 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' $\boldsymbol{\alpha}_+ = (0, 0, \ldots, 0, 0.5)$, $\boldsymbol{\alpha}_+$ will be a nonfit to ζ_n if and only if there exists a hyperplane V in the $\mathbf{x}$-space, which completely separates the data points from the two classes. Hence, the regression depth of the 'fit' $\boldsymbol{\alpha}_+$ can be viewed as the minimum number of misclassifications that can be achieved by a separating hyperplane V in the $\mathbf{x}$-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' $\boldsymbol{\alpha}_+$ 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

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

Here, the minimization of $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ w.r.t. $\boldsymbol{\alpha}$ and β gives the estimates $\widehat{\boldsymbol{\alpha}}_R$ 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 $\{(\boldsymbol{\alpha}, \beta) : \|(\boldsymbol{\alpha}, \beta)\| = 1\}$. It is also straight forward to verify that the minimization of $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ 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 $\mathbf{x}$-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 $\boldsymbol{\alpha}_+$. If the observations from the two classes are completely separable, it is fairly easy to see that $(\widehat{\boldsymbol{\alpha}}_R, \hat{\beta}_R)$ is a minimizer of $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ if and only if $\widehat{\boldsymbol{\alpha}}_R$ maximizes $U_{\mathbf{n}}(\boldsymbol{\alpha})$, and hence this $\widehat{\boldsymbol{\alpha}}_R$ is also an $\widehat{\boldsymbol{\alpha}}_H$.

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 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 $\mathbf{z}_i = (f_1(\mathbf{x}_i), f_2(\mathbf{x}_i), \ldots, f_m(\mathbf{x}))$, and perform a linear classification on that m-dimensional space. For instance, if we project the observations to the space of quadratic functions, it can be viewed as a linear classification with $m = d + \binom{d}{2}$ measurement variables, which eventually give rise to a quadratic separation in the original d-dimensional space. The quantities $U_{\mathbf{n}}(\boldsymbol{\alpha})$ and $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ can be optimized as before to get appropriate estimates of $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha} \in R^m$) and β, which are to be used to form the discriminating surfaces 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 ouliers 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 $\boldsymbol{\mu}_1 = (0,0)$ and $\boldsymbol{\mu}_2 = (2,2)$, and they have a common dispersion matrix $\boldsymbol{\Sigma} = \mathbf{I}_2$. 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 2.1(a)). But the scenario gets completely changed when five of the class-1 observations get replaced by outliers generated 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 2.1(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 avaialable diagnostic tool.

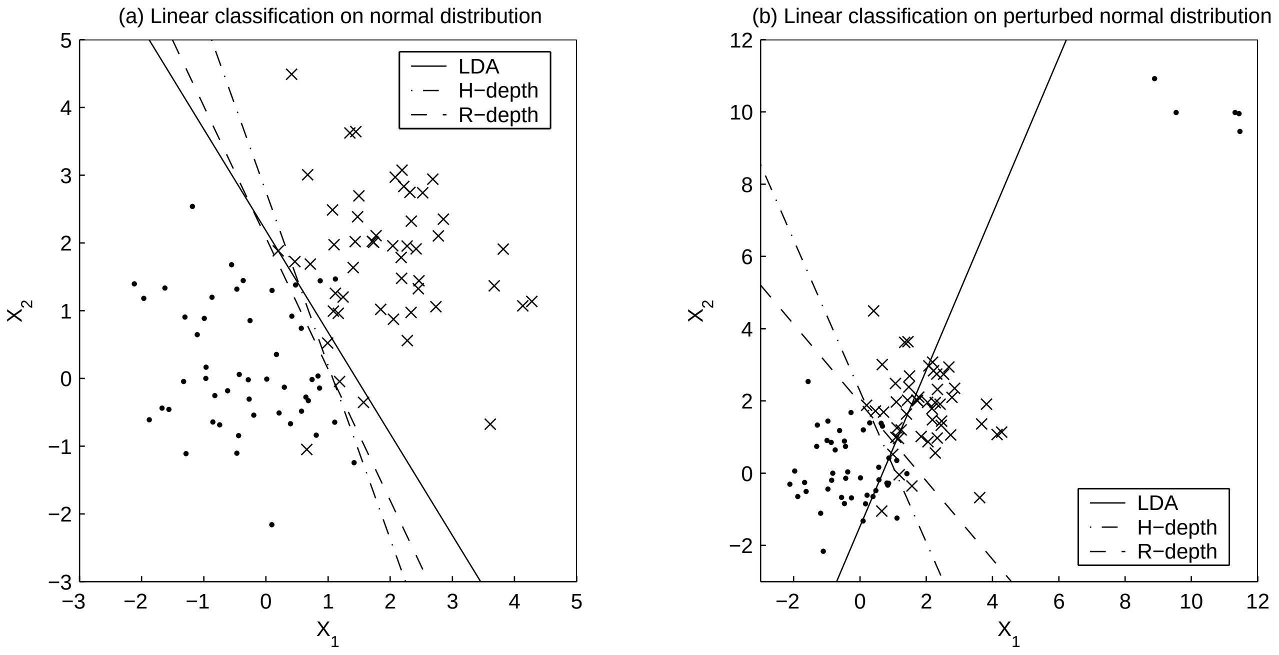

Figure 2.1 : Different linear classifiers for normal and perturbed normal distributions

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 m-dimensional space as described in Section 2.3 above, and $\widehat{\boldsymbol{\alpha}}_H$ is a maximizer of $U_{\mathbf{n}}(\boldsymbol{\alpha})$ while $\widehat{\boldsymbol{\alpha}}_R$, $\hat{\beta}_R$ are minimizers of $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ as before.

Theorem 3.1 : *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) $|U_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_H) - \max_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha})| \xrightarrow{a.s.} 0$ *as well as* $|U(\widehat{\boldsymbol{\alpha}}_H) - \max_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha})| \xrightarrow{a.s.} 0$, *and*

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

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,* $\widehat{\boldsymbol{\alpha}}_H$ *converges to* $\boldsymbol{\alpha}_H^*$ *and* $(\widehat{\boldsymbol{\alpha}}_R, \hat{\beta}_R)$ *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 $\Delta(\widehat{\boldsymbol{\alpha}}_R, \widehat{\beta}_R)$ can be viewed as the conditional average misclassification probability given the training sample, when the surface $\widehat{\boldsymbol{\alpha}}_R'\mathbf{z} + \widehat{\beta}_R = 0$ 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 Appendix. We state below some interesting and useful results for depth based linear and nonlinear classifiers that follow from this theorem.

Corollary 3.1 : *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 3.2 : *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 $f_i(\mathbf{x}) = g(\mathbf{x}-\boldsymbol{\mu}_i)$ $(i = 1, 2)$ for some location parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ 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 3.1, the average misclassification probability 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 $U(\boldsymbol{\alpha})$ 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 3.3 : *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

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

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the location and the scatter parameters, $\nu > 0$ and $\theta > d/2$ (see e.g., Fang, Kotz and Ng, 1989). When $\theta = (\nu + d)/2$ and ν is an integer, the corresponding distribution is known as the multivariate t-distribution with ν df. 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 m 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, 1998) or neural nets (see e.g., Ripley, 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 Appendix that if m grows with N in such a way that for all positive values of c, we have $\sum_{N \geq 1} N^{2m} e^{-cN} < \infty$, the convergence results in (i) and (ii) in Theorem 3.1 hold good. For instance, if m grows at the rate of N^{ρ} for any $0 < \rho < 1$, these convergence results remain valid.

4 Data analytic implementation

As we have already observed in Section 2, maximization of $U_{\mathbf{n}}(\boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\alpha}$ requires the computation of the half-space depth of the origin with respect to the data cloud formed by the m-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 $O(n_{\circ}^{2m})$ where $n_{\circ} = max\{n_1, n_2\}$. An algorithm due to Rousseeuw and Ruts (1996) can reduce the computational complexity to order $O(n_{\circ}^{2(m-1)} log\ n_{\circ})$. Similarly, maximization of $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ w.r.t. $\boldsymbol{\alpha}$ and β has computational complexity $O({n_{\circ}}^{m} log\ n_{\circ})$. 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).

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

Recall from Sections 2.1 and 2.2 that the maximization of $U_{\mathbf{n}}(\boldsymbol{\alpha})$ can be restricted to the $\boldsymbol{\alpha}$'s with $\|\boldsymbol{\alpha}\| = 1$ and the minimization of $\Delta(\boldsymbol{\alpha}, \beta)$ can be restricted to $(\boldsymbol{\alpha}, \beta)$'s with $\|(\boldsymbol{\alpha}, \beta)\| = 1$. However, since the order of the computational complexity increases rapidly with the dimension m, exact optimization of $U_{\mathbf{n}}(\boldsymbol{\alpha})$ and $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ is not feasible for high dimensional problems, and there one can only resort to some approximate optimization. In this article, we have used a procedure, where the indicator functions appearing in the expressions for $U_{\mathbf{n}}$ and $\Delta_{\mathbf{n}}$ 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 errors 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 $U_{\mathbf{n}}(\boldsymbol{\alpha})$ and $\Delta_{\mathbf{n}}(\boldsymbol{\alpha}, \beta)$ 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 classifiers based on half-space depth, after estimating $\boldsymbol{\alpha}$, we need to estimate β from the training sample. This is done by enumerating the order statistics of the projected data points $\widehat{\boldsymbol{\alpha}}_H^{'}\mathbf{z}_{1i}$ and $\widehat{\boldsymbol{\alpha}}_H^{'}\mathbf{z}_{2j}$ $(1 \leq i \leq n_1, 1 \leq j \leq n_2)$ along the estimated direction $\widehat{\boldsymbol{\alpha}}_H$. Fortunately, since we use linear projections, the computational complexity in obtaining the estimate $\widehat{\beta}_H$ does not increase with the dimension m.

4.2 Generalization of the procedure for multi-class problems

In a k-class $(k > 2)$ 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 $\binom{k}{2}$ pairs of classes, and then an observation is assigned to the population which has the maximum number of votes. However, 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 $\mathbf{x}$ is classified depending on the sign of $\boldsymbol{\alpha}^{'}\mathbf{z} + \beta$. So, if g is some monotonically increasing function on the real line satisfying $0 \leq g(x) \leq 1$, $g(0) = 0.5$ and $g(-x) = 1 - g(x)$ for every $x \in R$, we can use $g(\boldsymbol{\alpha}^{'}\mathbf{z} + \beta)$ 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, $1 - g(\boldsymbol{\alpha}^{'}\mathbf{z} + \beta)$ 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 k populations, and the observation can be classified to the population having highest combined posterior weight. However, we have applied pairwise coupling only for those rare observations, which did not get classified uniquely by the method of majority voting. In all our numerical studies reported in the following two sections, for coupling we have taken g to be the simple logistic function i.e., $g(x) = 1/(1+e^{-x})$. This choice is subjective and possibly many other choices will lead to similar results. Note that the logistic function used in approximate computation of depth as described in Section 4.1 has nothing to do with the choice of $g(x)$ here.

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 5.1A and 5.1B. 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 discussion.

As 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 (henceforth called LDA) performed very well as one would expect. However, the depth based methods also had a descent and comparable performance. But, in the case of multivariate Cauchy distribution, the depth based classifiers clearly outperformed LDA, and their performance was 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 $N(10\boldsymbol{\mu}_2, \mathbf{I})$ 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.

		Bayes			H-depth		R-depth	
		risk	n	LDA	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 5.1A : 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 5.1B : Results on linear discrimination : average misclassification rates (in percentages) with standard errors (dimensions 3 and 4).

Results obtained in the case of quadratic discrimination are reported in Table 5.2, and here too we found 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 $\mathbf{\Sigma_1} = \mathbf{I}$ and $\mathbf{\Sigma_2} = \mathbf{4I}$. The traditional quadratic discriminant analysis (henceforth called 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 descent 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.

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

6 Results from the analysis of benchmark data sets

We will now investigate the performance of the depth based classifiers on six well known data sets all of which except the first one are available from http://www.statlib.cmu.edu. In the case of the first two data sets (the vowel data 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 training and 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 have been reported in Table 6.1. In all the examples, sample proportions for different classes have been used as their prior probabilities.

6.1 Vowel data

We begin with a fairly well known data set related to a vowel recognition problem, where there are two measurement variables for each observation coming from one of the 10 classes. This data was created by Peterson and Barney (1952) by a spectrographic analysis of vowels on words formed by 'h' followed by a vowel and then followed by 'd'. There were 67 persons, who spoke the words, and the first two format frequencies (the two lowest frequencies of a speaker's vocal track) for 10 vowels were split into a training set consisting of 338 cases and a test set consisting of 333 observations. A scatter plot of this data set is given in Figure 6.1. This figure shows some significant overlaps among the competing classes, and this makes the data set a challenging one for any classification procedure.

For this data set, traditional LDA led to a test set error rate of 25.26 % (with a standard error of 0.02%), 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 0.02%) and 19.83% (with a standard error of 0.02%) respectively. Interestingly, as reported in Table 6.1, 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

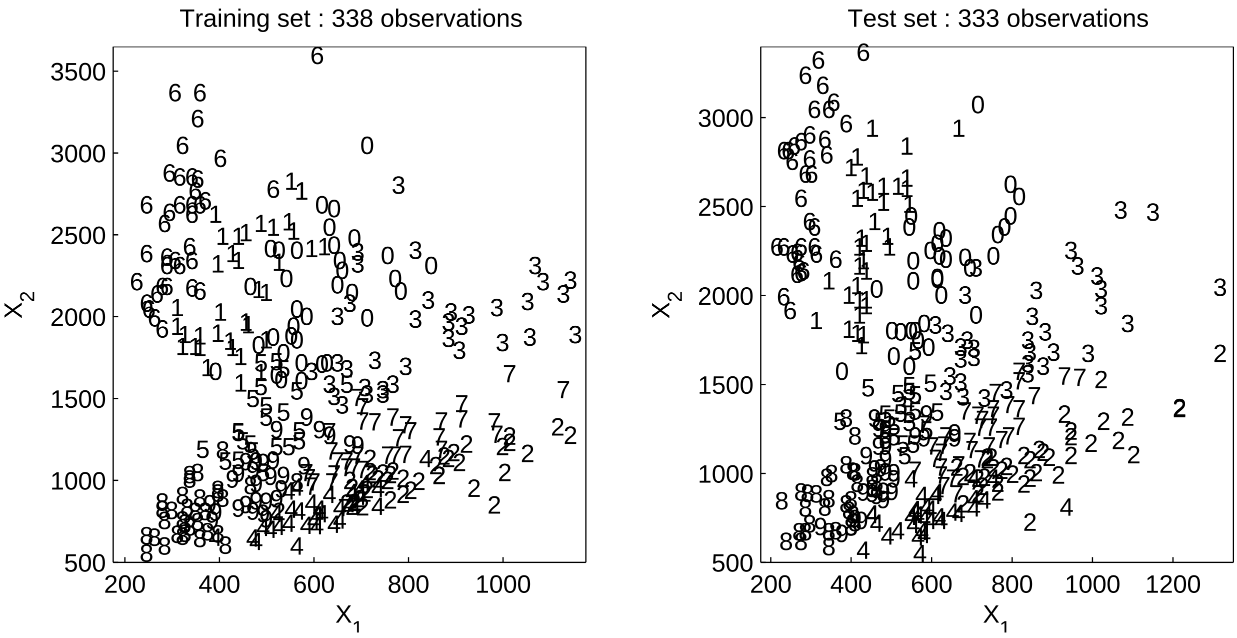

Figure 6.1 : Scatter plots for vowel data

fairly similar for this data.

6.2 Synthetic data

This bivariate data set has been used by Ripley (1994). It consists of bivariate observations from two competing populations. Here both the populations are bimodal in nature and they are equal mixtures of bivariate normal populations, which differ only in their location parameters. In this data set, the sizes of the training and the test sets are 250 and 1000 respectively. We report in Table 6.1 the average misclassification rates obtained for different methods applied on the test set. Here, we found that in linear as well as quadratic discriminant analysis, the error rates of the traditional and the depth based methods were fairly similar. Figure 6.2 shows 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.

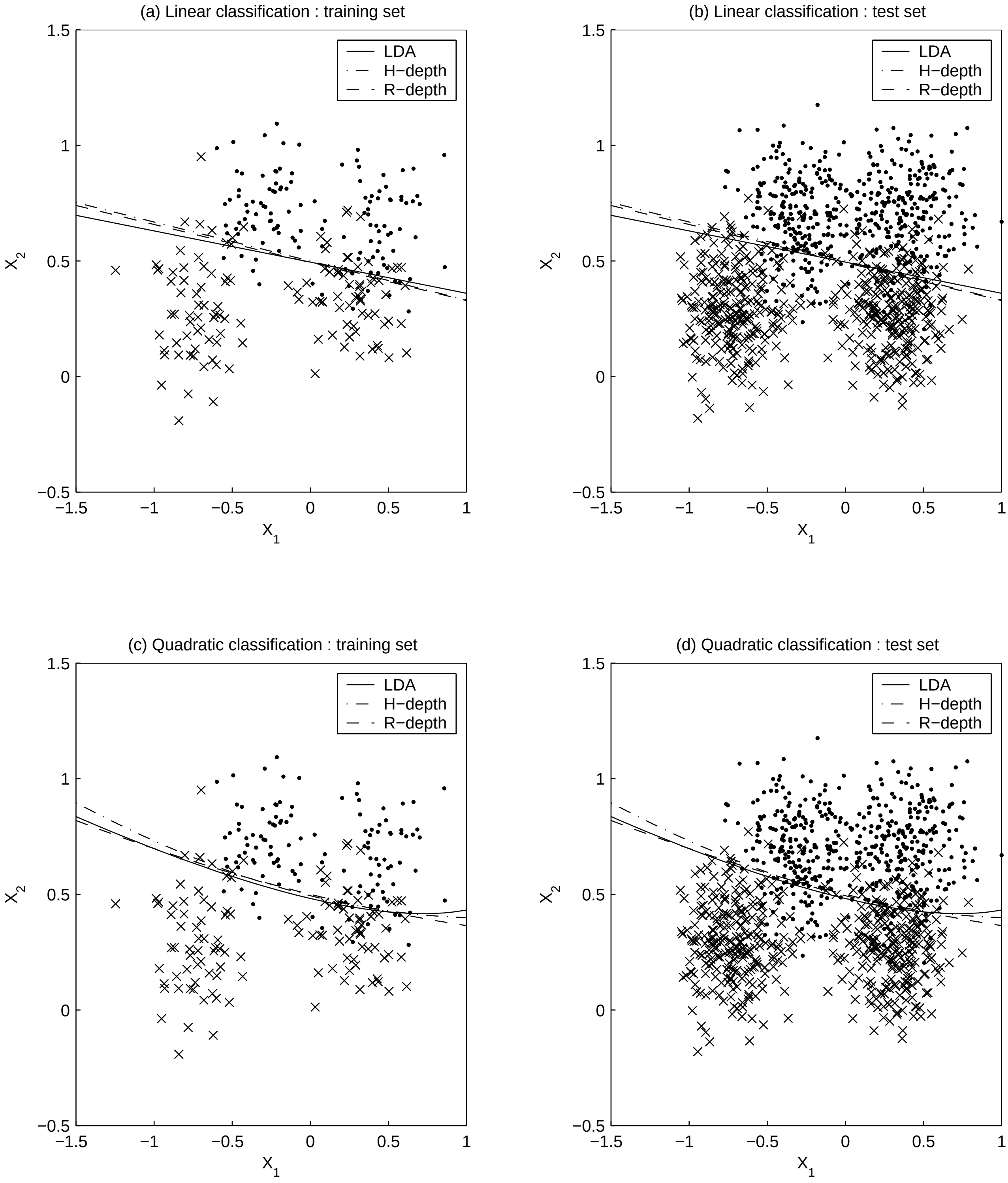

Figure 6.2 : Different linear and quadratic classifiers for synthetic data

6.3 Diabetes data

This data set contains measurements on five variables (fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight) and three classes ("overt diabetic", "chemical diabetic" and "normal") as reported in Reven and Miller (1979). There are 145 individuals with 33, 36 and 76 in these three classes according to some clinical classification. Unlike the vowel data and the synthetic data, this data set does not have any separate training and test sets, we formed these sets by randomly partitioning the data. We formed training samples of size 100 taking 25 observations from each of the first two populations and 50 observations from the third. The rest of the observations were used to form the corresponding test sets.

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% with a standard error (S.E.) of 0.07%, the H-depth and the R-depth based linear classifiers could reduce this error rate to 5.49% (S.E.= 0.06%) and 6.12% (S.E.= 0.06%) respectively. Though traditional QDA (error rate = 9.32%, S.E.= 0.06%) performed better than traditional LDA, it is quite apparent from the figures in Table 6.1 that depth based quadratic classifiers clearly outperformed the traditional one.

6.4 Bio-medical data

This data set was generated by Larry Cox and used by Cox, Johnson and Kafadar (1982) in the annual meeting on "Exposition of Statistical Graphics Technology". This data set contains the information on four different measurements on each of the 209 blood samples (134 for "normals" and 75 for "carriers"). Out of the 209 observations, 15 have missing values, and we have removed those observations and applied the classification methods on the remaining 194 cases (127 for "normals" and 67 for "carriers"). 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 as the corresponding test cases.

Here also the depth based linear classifiers outperformed traditional LDA. As shown in Table 6.1, LDA had an error rate of 15.96% (S.E.= 0.07%), while the H-depth and the R-depth based classifiers could reduce it upto 10.87% (S.E.= 0.07%) and 11.03% (S.E.= 0.07%) respectively. Figures reported in Table 6.1 also indicate that even in the case of quadratic discriminant analysis, these depth based classifiers have a slight edge over traditional QDA for this data set.

6.5 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 6.1 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.

6.6 Iris data

As the last example of this section, we consider the famous Iris data (Fisher, 1936), which contains the measurements on four different features (sepal length, sepal width, petal length and petal width) on each of the 150 observations coming from three different types of Iris plants : 'Iris Setosa', 'Iris Virginica' and 'Iris Versicolor'. 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 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 6.1 show that both the linear and the quadratic versions of the depth based methods could produce a descent and comparable performance on 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 6.1 : Results on benchmark data sets : average misclassification rates (in percentages) with standard errors.

7 Concluding remarks

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). Along with L_1 depth, Ghosh and Chaudhuri (2004) used other depth functions to construct their maximum depth classifiers. However, for classifying a new observation, these classifiers need to calculate its depths with respect to different competing populations, and for that the full training sample have to be stored. Moreover, it is difficult to generalize these classifiers for unequal prior cases (see Ghosh and Chaudhuri, 2004). On the other hand, the depth based classifiers proposed in this article require less storage and computing time for classifying future observations, and at the same time it provides a good lower dimensional view of class separabiltiy.

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, 1996) and support vector machines (see e.g., Vapnik, 1998 – henceforth we will write SVM as an abbreviation for this) 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 7.1). 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.

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

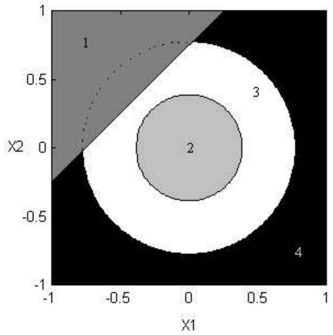

Figure 7.1 : A four class problem for comparing different classifiers

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.

8 Appendix : Proofs

In order to prove Theorem 3.1, we will need the following result, which follows directly from the proof of Lemma A of Serfling (1980), p. 200.

Result 8.1 : 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 3.1 : (i) $U_{\mathbf{n}}(\boldsymbol{\alpha})$ is a generalized U-statistic (see e.g., Serfling, 1980) having the bounded kernel function $h(\boldsymbol{\alpha}'\mathbf{z}_1, \boldsymbol{\alpha}'\mathbf{z}_2) = I\{\boldsymbol{\alpha}'\mathbf{z}_1 > \boldsymbol{\alpha}'\mathbf{z}_2\}$ $(0 \leq h \leq 1)$. Now, without loss of generality, let us assume that $n_1 \leq n_2$ and define

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

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, $U_{\mathbf{n}}(\boldsymbol{\alpha})$ can be expressed as

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

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

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

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

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

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

Now, it is quite easy to see that

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

using Result 8.1, we get for any $t > 0$,

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

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

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

Now, the set of hyperplanes in $V = \{\mathbf{y} : \boldsymbol{\alpha}^{'}\mathbf{y} = 0\}$ in R^m, which pass through the origin has VC dimension m (see e.g., Pollard, 1984; Vapnik, 1998). So, the sets of the form $\{\mathbf{y} : \boldsymbol{\alpha}^{'}\mathbf{y} > 0\}$ has a polynomial discrimination with m 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

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

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

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

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

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

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

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

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

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

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

Then, using the fact that $\sum_{N\geq 1} N^{m+1}e^{-cN} < \infty$ for any $c > 0$, it follows from Borel Cantelli lemma that $\sup_{\boldsymbol{\alpha},\beta} |\Delta_{\mathbf{n}}(\boldsymbol{\alpha},\beta) - \Delta(\boldsymbol{\alpha},\beta)| \to 0$ 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 $|\Delta(\widehat{\boldsymbol{\alpha}}_R, \hat{\beta}_R) - \min_{\boldsymbol{\alpha},\beta} \Delta(\boldsymbol{\alpha},\beta)| \to 0$ and $|\Delta_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_R, \hat{\beta}_R) - \min_{\boldsymbol{\alpha},\beta} \Delta(\boldsymbol{\alpha},\beta)| \to 0$ almost surely as $N \to \infty$.

Let us next assume that the maximizer $\boldsymbol{\alpha}_H^*$ of $U(\boldsymbol{\alpha})$ is unique. We have already shown $U(\widehat{\boldsymbol{\alpha}}_H)$ converges to $U(\boldsymbol{\alpha}_H^*)$ as $N \to \infty$ on a set of probability one. Consequently, on the same set, if $\widehat{\boldsymbol{\alpha}}_H$ converges, it has to converge to $\boldsymbol{\alpha}_H^*$ in view of the uniqueness of $\boldsymbol{\alpha}_H^*$ and the continuity of the function $U(\boldsymbol{\alpha})$. Since $\widehat{\boldsymbol{\alpha}}_H$ always lies in the compact surface of the unit ball in R^m (see Sections 2.1 and 4.1), any subsequence of the sequence of this estimate will have a further convergent subsequence converging to $\boldsymbol{\alpha}_H^*$ on that set of probability one. Hence, $\widehat{\boldsymbol{\alpha}}_H$ must converge to $\boldsymbol{\alpha}_H^*$ almost surely.

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

Proof of Corollary 3.1 : In Theorem 3.1, we have proved that $|\Delta(\widehat{\boldsymbol{\alpha}}_R, \hat{\beta}_R) - \min_{\boldsymbol{\alpha},\beta} \Delta(\boldsymbol{\alpha}, \beta)| \to 0$ almost surely as $N \to \infty$. Note that $\Delta(\widehat{\boldsymbol{\alpha}}_R, \hat{\beta}_R)$ is the conditional average misclassification probability for a future observation given the current training sample. Taking expectation of $\Delta(\widehat{\boldsymbol{\alpha}}_R, \hat{\beta}_R)$ 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 8.1 : 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 $f_i(\mathbf{x}) = g(\mathbf{x} - \boldsymbol{\mu}_i)$ $(i = 1, 2)$ for some location parameters $\boldsymbol{\mu}_i$ 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) $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is a maximizer of $U(\boldsymbol{\alpha})$ as well as a minimizer $\Delta(\boldsymbol{\alpha}, \beta)$ for a proper choice of β.

Proof of Lemma 8.1 : (i) Because of elliptic symmetry with location shift, the density functions f_1 and f_2 can be expressed as

$$f_1(\mathbf{x}) = C_d|\Sigma|^{-1/2} h\{(\mathbf{x} - \boldsymbol{\mu}_1)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1)\} \text{ and } f_1(\mathbf{x}) = C_d|\Sigma|^{-1/2} h\{(\mathbf{x} - \boldsymbol{\mu}_2)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)\},$$

where C_d is a constant (depends on dimension d) and h 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

$$\begin{aligned} f_1(\mathbf{x}) \geq f_2(\mathbf{x}) \quad &\Leftrightarrow (\mathbf{x} - \boldsymbol{\mu}_1)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) \leq (\mathbf{x} - \boldsymbol{\mu}_2)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \\ &\Leftrightarrow (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} \geq \tfrac{1}{2} \left[\boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \right]. \end{aligned}$$

This proves that an optimal linear classifier is a Bayes classifier and $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is a minimizer of $\Delta(\boldsymbol{\alpha}, \beta)$ with a proper choice of β.

(ii) As the distributions have a common elliptically symmetric form with location parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and common scatter matrix Σ, their characteristic functions are of the form

$\phi_{f_1}(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu}_1}\psi(\mathbf{t}'\Sigma\mathbf{t})$ and $\phi_{f_2}(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu}_2}\psi(\mathbf{t}'\Sigma\mathbf{t})$ for some common scalar function ψ.

Now define $Y = \boldsymbol{\alpha}'\{(\mathbf{X}_1 - \mathbf{X}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}/(\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha})^{1/2}$, 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 $\phi_Y(t) = \{\psi(t^2)\}^2$. Clearly, the distribution of Y is symmetric about 0, and it is free from population parameters like the $\boldsymbol{\mu}$'s and the $\boldsymbol{\Sigma}$. Therefore, $P\{\boldsymbol{\alpha}'(\mathbf{X}_1 - \mathbf{X}_2) > 0\}$ can be expressed as

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

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

Proof of Corollary 3.2 : Lemma 8.1 implies that, under the given conditions, the linear classifier with $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\beta = (\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1)/2$ is a Bayes classifier. Consequently, it follows from Corollary 3.1 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 3.1 that the regression depth based linear classifier itself converges almost surely to that Bayes classifier.

When $U(\boldsymbol{\alpha})$ has a unique maximizer $\boldsymbol{\alpha}_H^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ (e.g., when the distribution function F_Y in the proof of Lemma 8.1 is strictly increasing), it follows from Theorem 3.1 that $\widehat{\boldsymbol{\alpha}}_H$ converges almost surely to $\boldsymbol{\alpha}^*$ 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 $Y_{1,\mathbf{n}} = \widehat{\boldsymbol{\alpha}}_H'\mathbf{X}_1$, $Y_{2,\mathbf{n}} = \widehat{\boldsymbol{\alpha}}_H'\mathbf{X}_2$, $Y_1 = \boldsymbol{\alpha}^{*'}\mathbf{X}_1$ and $Y_2 = \boldsymbol{\alpha}^{*'}\mathbf{X}_2$. Then, in view of almost sure convergence of $\widehat{\boldsymbol{\alpha}}_H$ 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 $\sup_{\beta} |\Delta(\widehat{\boldsymbol{\alpha}}_H, \beta) - \boldsymbol{\Delta}(\boldsymbol{\alpha}^*, \beta)| \to 0$ almost surely as $N \to \infty$.

On the other hand, from the proof of (ii) in Theorem 3.1, it is quite clear that $\sup_{\beta} |\Delta_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_H, \beta) - \Delta(\widehat{\boldsymbol{\alpha}}_H, \beta)| \to 0$ almost surely as $N \to \infty$. Hence, $\sup_{\beta} |\Delta_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_H, \beta) - \Delta(\boldsymbol{\alpha}^*, \beta)| \to 0$ almost surely as $N \to \infty$.

It now follows from arguments similar to those used in the proof of Theorem 3.1, $|\Delta_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H) - \min_{\beta} \Delta(\boldsymbol{\alpha}^*, \beta)| = |\Delta_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H) - \min_{\boldsymbol{\alpha}, \beta} \Delta(\boldsymbol{\alpha}, \beta)| \to 0$ almost surely as $N \to \infty$. Also, we must have $|\Delta_{\mathbf{n}}(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H) - \Delta(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H)| \to 0$ almost surely as $N \to \infty$. Hence, $\Delta(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H)$ converges almost surely to $\min_{\boldsymbol{\alpha}, \beta} \Delta(\boldsymbol{\alpha}, \beta)$, which is the Bayes risk in this case.

Once again, note that $\Delta(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H)$ is the conditional average misclassification probability for a future observation given the current training sample. Taking expectation of $\Delta(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H)$ 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 $\hat{\beta}_H$ 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 $\Delta(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H)$ converges to the Bayes risk and $\widehat{\boldsymbol{\alpha}}_H$ converges to $\boldsymbol{\alpha}^*$ as $N \to \infty$ on a set with probability one. So, on this set $\hat{\beta}_H$ must remain bounded as otherwise in view of the convergence of $\widehat{\boldsymbol{\alpha}}_H$ to $\boldsymbol{\alpha}^*$, $\Delta(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H)$ will converge to 0.5 along a subsequence along which $|\hat{\beta}_H| \to \infty$ as $N \to \infty$. On the other hand, whenever $\hat{\beta}_H$ converges to a real number β (say), in view of the continuity of Δ, $\Delta(\widehat{\boldsymbol{\alpha}}_H, \hat{\beta}_H)$ must converge to $\Delta(\alpha^*, \beta)$ on that set of probability one. Since any bounded sequence must have a convergent subsequence, it is now obvious that $\hat{\beta}_H$ must converge to β^*, where $\Delta(\alpha^*, \beta^*) = \min_{\boldsymbol{\alpha}, \beta} \Delta(\boldsymbol{\alpha}, \beta)$, 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 $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$,

$$\begin{aligned}
\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x}) \quad & \Leftrightarrow \pi_1 \ |\boldsymbol{\Sigma}|^{-1/2} \ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)} > \pi_2 \ |\boldsymbol{\Sigma}|^{-1/2} \ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)} \\
& \Leftrightarrow (\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - (\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) > C, \text{ where } \ C = 2 \ log\,(\pi_2/\pi_1)\,. \\
& \Leftrightarrow 2\mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \{\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\} + C.
\end{aligned}$$

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 3.3 : 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 $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$,

$$\begin{aligned}
\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x}) \quad & \Leftrightarrow \pi_1 \ |\boldsymbol{\Sigma}_1|^{-1/2} \ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)} > \pi_2 \ |\boldsymbol{\Sigma}_2|^{-1/2} \ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)} \\
& \Leftrightarrow (\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - (\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) > C,
\end{aligned}$$

where $C = 2 \ log \left(\frac{\pi_2|\boldsymbol{\Sigma}_1|^{1/2}}{\pi_1|\boldsymbol{\Sigma}_2|^{1/2}}\right)$. 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

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

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

$$\begin{aligned}
& \Leftrightarrow \pi_1|\boldsymbol{\Sigma}_1|^{-1/2}\{1 + \nu^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\}^{-\theta} > \pi_2|\boldsymbol{\Sigma}_2|^{-1/2}\{1 + \nu^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\}^{-\theta} \\
& \Leftrightarrow \left\{\frac{1 + \nu^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}{1 + \nu^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}\right\}^{-\theta} > K \ \text{ for } K = \frac{\pi_2|\boldsymbol{\Sigma}_2|^{-1/2}}{\pi_1|\boldsymbol{\Sigma}_1|^{-1/2}} \\
& \Leftrightarrow \left\{\frac{\nu + (\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}{\nu + (\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}\right\} < C = K^{-1/\theta} \\
& \Leftrightarrow (\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - C(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - (C-1)\nu < 0.
\end{aligned}$$

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.

References

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

[2] Bai, Z.-D. and He, X. (1999) Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *Ann. Statist.*, **27**, 1616-1637.

[3] 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.

[4] Chaudhuri, P. and Sengupta, D. (1993) Sign tests in multi-dimension : inference based on the geometry of the data cloud. *J. Amer. Statist. Assoc.*, **88**, 1363-1370.

[5] Christmann, A. and Rousseeuw, P. (2001) Measuring overlap in binary regression. *Comput. Statist. Data Analysis*, **37**, 65-75.

[6] Christmann, A. (2002) Classification based on support vector machine and on regression depth. *Statistics in Industry and Technology : Statistical Data Analysis*, 341-352.

[7] 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. *Comput. Statist.*, **17**, 273-287.

[8] Cox, L. H., Johnson, M. M. and Kafadar, K. (1982) Exposition of statistical graphics technology, *ASA Proc. Statist. Comput. Section*, pp. 55-56.

[9] Donoho, D. and Gasko, M. (1992) Breakdown properties of location estimates based on half-space depth and projected outlyingness. *Ann. Statist.*, **20**, 1803-1827.

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

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

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

[13] Friedman, J. H. (1989) Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, **84**, 165-175.

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

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

[16] Hand, D. J. (1981) *Discrimination and Classification*, Wiley, New york.

[17] Ghosh, A. K. and Chaudhuri, P. (2004) On maximum depth classifiers. *Submitted for publication.*

[18] 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.

[19] Jornsten, R. (2004) Clustering and classification based on L_1 data depth. To appear in *J. Multivariate Analysis.*

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

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

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

[23] He, X. and Wang, G. (1997) Convergence of depth contours for multivariate datasets. *Ann. Statist.*, **25**, 495-504.

[24] Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13-30.

[25] Liu, R. (1990) On notion of data depth based on random simplicies. *Ann. Statist.*, **18**, 405-414.

[26] Liu, R., Parelius, J. and Singh, K. (1999) Multivariate analysis of the data-depth : descriptive statistics and inference. *Ann. Statist.*, **27**, 783-858.

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

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

[29] Nolan, D. (1992) Asymptotics for multivariate trimming. *Stoc. Proc. Appl.*, **42**, 157-169.

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

[31] Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer Verlag, New York.

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

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

[34] 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.

[35] Rousseeuw, P.J. and Ruts, I. (1996) Algorithm AS 307: bivariate location depth. *Applied Statistics* , **45**, 516-526.

[36] Rousseeuw, P.J. and Struyf, A. (1998) Computing location depth and regression depth in higher dimensions. *Statistics and Computing* , **8**, 193-203.

[37] Rousseeuw, P.J. and Hubert, M. (1999) Regression depth (with discussions). *J. Amer. Statist. Assoc.* , **94**, 388-402.

[38] 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.

[39] Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

[40] 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.

[41] Tukey, J. (1975) Mathematics and the picturing of data. *Proc. Inter. Cong. Math., Vancouver*, 523-531.

[42] Vapnik, V. N. and Chervonenkis, A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theo. Prob. Appl.*, **16**, 264-281.

[43] Vapnik, V. N. (1998) *Statistical Learning Theory.* Wiley, New York.

[44] 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.

[45] Zhu, M. and Hastie, T. (2003) Feature extraction for nonparametric discriminant analysis. *J. Coumput. Graph. Statist.*, **12**, 101-120.

[46] Zuo, Y. and Serfling, R. (2000a) General notions of statistical depth functions. *Ann. Statist.*, **28**, 461-482.

[47] Zuo, Y. and Serfling, R. (2000b) Structural properties and convergence results for contours of sample statistical depth functions. *Ann. Statist.*, **28**, 483-499.