On maximum depth and related classifiers

Anil K. 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

Over the last couple of decades, data depth has emerged as a powerful exploratory and inferential tool for multivariate data analysis with wide-spread applications. This paper investigates possible use of different notions of data depth in nonparametric discriminant analysis. First, we consider the situation where the prior probabilities of the competing populations are all equal and investigate classifiers that assign an observation to the population with respect to which it has the maximum location depth. We propose a different depth based classification technique for unequal prior problems, which is also useful for equal prior cases, especially when the populations have different scatters and shapes. We use some simulated data sets as well as some benchmark real examples to evaluate the performance of these depth based classifiers. Large sample behavior of the misclassification rates of these depth based nonparametric classifiers have been derived under appropriate regularity conditions.

Keywords and Phrases : Bayes risk, cross-validation, data depth, elliptic symmetry, kernel density estimation, location shift models, Mahalanobis distance, misclassification rates, Vapnik Chervonenkis dimension.

1 Introduction : data depth and discriminant analysis

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 various features and properties of univariate distributions to multivariate distributions. The notions of multivariate median, multivariate L-statistics, tests for the center of elliptic symmetry, measures of multivariate dispersion and skewness are some well known examples of its application (see e.g., Chaudhuri and Sengupta, 1993; Liu and Singh, 1993; Liu, Parelius and Singh, 1999; Vardi and Zhang, 2000; Mosler, 2002). Several notions of depth functions are available in literature. Some of these depth functions are briefly described below.

- Mahalanobis depth (MD) (see e.g., Mahalanobis, 1936; Liu and Singh, 1993) of an observation $\mathbf{x}$ w.r.t. the distribution F is defined to be

$$MD(F, \mathbf{x}) = \left\{1 + (\mathbf{x} - \boldsymbol{\mu}_F)' \boldsymbol{\Sigma}_F^{-1} (\mathbf{x} - \boldsymbol{\mu}_F)\right\}^{-1},$$

  where $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ are the mean vector and the dispersion matrix of the distribution F.

- Half-space depth (HD) (see e.g., Tukey, 1975) of $\mathbf{x}$ w.r.t. F is defined as the minimum probability measure of any closed half-space containing $\mathbf{x}$.

$$HD(F, \mathbf{x}) = \inf_{H} \left\{ P_F(H) : H \text{ is a closed half space in } R^d, \text{ and } \mathbf{x} \in H \right\}.$$

- Simplicial depth (SD) (see e.g., Liu, 1990) of $\mathbf{x}$ w.r.t. F is defined to be the probability that $\mathbf{x}$ belongs to a random simplex in R^d.

$$SD(F, \mathbf{x}) = P_F \left\{ \mathbf{x} \in S[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{d+1}] \right\},$$

  where $S[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{d+1}]$ is a d-dimensional simplex formed by $(d+1)$ i.i.d. observations $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_{d+1}$ from F.

- Majority depth (MJD) (see e.g., Singh, 1991; Liu and Singh, 1993) of $\mathbf{x}$ w.r.t. F is defined as the probability that $\mathbf{x}$ belongs to the major side (i.e. the half-space with larger probability measure) of a random hyperplane passing through d data points in R^d.

$$MJD(F, \mathbf{x}) = P_F \left\{ \mathbf{x} \text{ belongs to the major side of } \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d \right\},$$

  where $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d$ are i.i.d. observations from F.

- Projection depth (PD) (see e.g., Stahel, 1981; Donoho, 1982) of $\mathbf{x}$ w.r.t. F is defined as the worst case outlyingness of $\mathbf{x}$ w.r.t. one dimensional median in any one-dimensional projection.

$$PD(F, \mathbf{x}) = \sup_{\|\boldsymbol{\alpha}\|=1} \left[\{\boldsymbol{\alpha}^{'}\mathbf{x} - Median(\boldsymbol{\alpha}^{'}\mathbf{X})\} / MAD(\boldsymbol{\alpha}^{'}\mathbf{X}) \right],$$

  where $MAD(\mathbf{Y}) = Median\left(|\mathbf{Y} - Median(\mathbf{Y})|\right)$ and $\mathbf{X} \sim F$.

- Simplicial volume depth (SVD) (Zuo and Serfling, 2000a, 2000b) is closely related to Oja median (Oja, 1983). SVD of an observation $\mathbf{x}$ w.r.t. F can be expressed as

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

  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 $\boldsymbol{\Sigma}_F$ is the scatter matrix of the distribution F. Note that the division by $|\boldsymbol{\Sigma}_F|^{1/2}$ is required only to make the depth function affine invariant like the other depth functions mentioned above.

- The notion of spatial depth (SPD) or L_1 depth (Vardi and Zhang, 2000; Serfling, 2002) follows the work of Chaudhuri (1996) and Kolchinskii (1997) on spatial quantiles. SPD of an observation $\mathbf{x}$ w.r.t. F is defined as

$$SPD(F, \mathbf{x}) = 1 - \left\| E_F \left\{ \frac{\mathbf{x} - \mathbf{X}}{\|\mathbf{x} - \mathbf{X}\|} \right\} \right\|, \quad \text{where } \mathbf{X} \sim F.$$

Spatial depth has some nice properties. When $d \geq 2$, for all F, $E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}$ is a continuous and "monotonic transformation" on R^d, and it uniquely determines the distribution function F (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 $\mathbf{0}$, and hence $SPD(F, \mathbf{x})$ is expected to attain

its maximum value 1. On the other hand, if the point moves away from the center, SPD approaches the value 0. Unlike other depth functions, SPD is easy to compute for high dimensional data, and one can define SPD even for infinite dimensional Hilbert spaces (see e.g., Chaudhuri, 1996). This depth function is invariant under rotation and if the same scale transformation is done on all co-ordinate variables. One can also make it affine invariant by taking $\mathbf{\Sigma}_F^{-1/2}(\mathbf{x}-\mathbf{X})$ instead of $\mathbf{x}-\mathbf{X}$.

Various other well known depth functions like likelihood depth, convex hull peeling depth and zonoid depth have been studied by Liu, Parelius and Singh (1999), Zuo and Serfling (2000a, 2000b), Mosler (2002) and Mizera (2002). Apart from likelihood depth, all these depth functions are affine invariant in nature. Likelihood depth also preserves the ordering of the depth functions under affine transformations.

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 in the literature (see e.g. Liu, 1990; Nolan, 1992; Donoho & Gasko, 1992; Liu & Singh, 1993; He and Wang, 1997; Zuo and Serfling, 2000a, 2000b). To make it notationally simpler, instead of $D(F_j,\mathbf{x})$ and $D(F_{n_j},\mathbf{x})$ we will write $D(j,\mathbf{x})$ and $D_n(j,\mathbf{x})$, respectively, to denote the theoretical (population) and the empirical (sample) depth of $\mathbf{x}$ with respect to the j^{th} population.

Like other useful applications of depth functions in multivariate statistics, different notions of data depth can also be used for the purpose of discriminant analysis, where the objective is to classify an observation into one of several competing populations. Given the prior probabilities π_j's and the density functions f_j's of these populations, the optimal Bayes classification rule assigns an observation $\mathbf{x}$ to the population having the maximum posterior probability at $\mathbf{x}$ (i.e., it assigns $\mathbf{x}$ to the i^{th} population, where $i=\arg\max \pi_j f_j(\mathbf{x})$). This classifier has the lowest possible average misclassification rate known as the optimal Bayes risk. However, in practice f_j's ($j=1,2,\ldots,J$) are unknown, and they have to be estimated using the available training sample. Parametric methods like Fisher's (see Fisher, 1936) linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are motivated by specific distributional assumptions on the competing populations, which may not be valid in practice. Further, these traditional classifiers use estimates of the unknown population parameters (e.g., means, variances and correlations) based on the moments of the training sample observations, and this makes these methods very sensitive to outliers and extreme values. Christmann and Rousseeuw (2001) and Christmann, Fischer and Joachims (2002) used the idea of regression depth (see e.g., Rousseeuw and Hubert, 1999) for classification between two competing populations. Ghosh and Chaudhuri (2004b) extended this idea for multi-class problems, where they used regression depth and half space depth (see e.g., Tukey, 1975) to construct linear and nonlinear discriminating surfaces. In these depth based methods, one assumes a finite dimensional parametric form (usually linear or quadratic) for the separating surface and uses the distributional geometry of the data cloud to estimate the associated parameters. In that sense, these classifiers though distribution free are not fully nonparametric but only semiparametric in nature. Performance of these depth based semiparametric classifiers and their asymptotic properties have been studied in Ghosh and Chaudhuri (2004b). The objective of this article is to demonstrate how various notions of data depth can also be used to develop fully nonparametric classifiers (see also Liu, 1990, pp. 408-409). In the next section, we investigate maximum depth classifiers for classification in

equal prior cases. Later, we will introduce another depth based classification technique, which works under a more general set up.

2 Maximum depth classifiers

Unlike the parametric and semiparametric classification methods, maximum depth classifiers do not assume any specific parametric form of the separating surface nor do they assume any particular type of probability distribution for the populations. Instead, they classify an observation to the class with respect to which it has the maximum location depth. These classifiers can be expressed as

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

where n_j is the number of training sample observations, $D_{n_j}(j, \mathbf{x})$ is the empirical depth of $\mathbf{x}$ in the j^{th} population, and the prior probabilities of the competing classes are assumed to be equal. It is straight forward to see that when MD is used, such a depth based classifier leads to a linear or a quadratic classifier depending on whether a common scatter matrix is used for all of the competing populations or not. So if one is interested in a more flexible procedure than just a classifier with a linear or quadratic class boundaries, MD will not be an appropriate choice. 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 $|\mathbf{\Sigma}|^{1/2}$ in the denominator of the expression of SVD^{δ} when it is used for maximum depth classification. Recently maximum depth classifiers based on SPD have been studied by Jornsten (2004) for classification of microarray gene expression data, where she used SPD also for cluster analysis. A cluster analysis method based on zonoid depth has also been investigated in Hoberg (2000).

2.1 Misclassification rates and asymptotic optimality

When the population distribution is elliptic with density function strictly decreasing in every direction from its center of symmetry, some of the population depth functions also satisfy that monotonicity property (see e.g., Zuo and Serfling, 2000a), and they turn out to be a decreasing function of the population Mahalanobis distance (Mahalanobis, 1936). HD, SD, MJD, MD, PD and SVD (for $\delta \geq 1$) are some of the depth functions with this property (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 population depth functions are equivalent to population Mahalanobis distance for classification purpose, and they all are equivalent to the optimal Bayes classifier. However, in practice, population depth functions are not available, and one has to use the empirical depth functions, which are natural estimates for their population counterparts, in order to classify an observation. The following theorems establish asymptotic optimality of maximum depth classifiers based on some of the empirical depth functions. From now on, we will assume that all the populations possess densities, which are continuous and positive over the entire d-dimensional space. Also, the average misclassification rate of an empirical depth based classifier will be given by

$$\Delta_{\mathbf{n}} = \sum_{j=1}^{J} \pi_j \; P\{\mathtt{d}_D(\mathbf{X}) \neq j \mid \mathbf{X} \in j^{th}\text{population}\} = \sum_{j=1}^{J} \; P\{\mathtt{d}_D(\mathbf{X}) \neq j \;\text{ and }\; \mathbf{X} \in j^{th}\text{population}\} \;,$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_J)$ is the vector of training sample sizes for different classes, and π_j is the prior probability for the j^{th} population ($j = 1, 2, \ldots, J$). Note that in the above definition of $\Delta_{\mathbf{n}}$, P denotes the unconditional joint probability involving both the probability distribution of the training sample as well as that of the test case $\mathbf{X}$.

Theorem 2.1 : *Suppose that the population density functions $f_1, f_2, \ldots, f_J$ are elliptically symmetric, and $f_j(\mathbf{x}) = g(\mathbf{x} - \boldsymbol{\mu}_j)$ for some location parameters $\boldsymbol{\mu}_j$ and a common density function g with $g(k\mathbf{x}) \leq g(\mathbf{x})$ for every $\mathbf{x}$ and $k > 1$. Now define $\mathbf{n} = (n_1, n_2, \ldots, n_J)$ and $\Delta_{\mathbf{n}}$ as above. 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 2.2 : *Assume the same set up as in Theorem 2.1. 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 2.3 : *Assume all the conditions of Theorem 2.1. 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 the optimal Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.*

3 Data analytic implementation of the classifiers

Among various notions of depth functions, MD is surely the simplest one to calculate, but as we have mentioned earlier, it can only lead to classification using linear or quadratic discriminant functions (LDA and QDA). Further, computational simplicity in the case of MD is a consequence of using sample moments that are always very easy to compute. When more robust estimates for the location vector and the scatter matrix are used as has been done by some recent authors (see e.g., He and Fung, 2000; Croux and Dehon, 2001; Hubert and Van Driessen, 2003), this computational simplicity is lost. Note that like MD, SVD^{δ} also depends on some empirical moments based on the training sample, and consequently they both are sensitive to outliers and extreme values. Many of the 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. Among such depth based classifiers, SPD in practice has some advantages. We have already pointed out that it is computationally less expensive than most of the other depth functions, and it can be used for classification even in infinite dimensional Hilbert spaces. Since the empirical version of SPD is continuous in $\mathbf{x}$ for $d \geq 2$, there is almost no possibility of ties, while ties may cause problems for depth functions like HD, SD and MJD because of their step function (piecewise constant) like nature.

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; Rousseeuw 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 algorithms. Such an approximate algorithm for HD was proposed in Ghosh and Chaudhuri (2004b). This approximation allows us to use derivatives of

certain smooth functions to find out the direction of steepest ascent or descent of the objective function to be optimized. In this paper, for all problems with $d > 2$, we have adopted this approximation for computing HD of an observation. Exact version of HD is used for bivariate data sets 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.

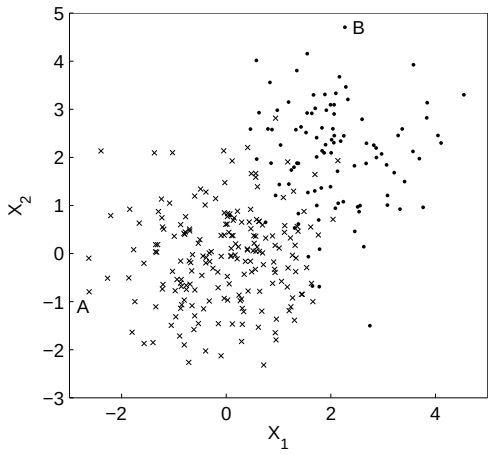

Figure 3.1 : Scatter plot for simulated data

Apart from computational difficulty, HD and SD have another problem in higher dimensions. Consider the following example of a two-class problem where the classes are bivariate normal with means $\boldsymbol{\mu}_1 = (0,0)$, $\boldsymbol{\mu}_2 = (2,2)$ and common dispersion matrix $\mathbf{I_2}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. We have generated 100 observations from each class, and the scatter plot of this data set is given in Figure 3.1. In this figure, one can notice some observations which have zero empirical depth for HD as well as SD with respect to both the classes. For instance, the observations 'A' and 'B' (see Figure 3.1) clearly belong to two different classes but both have zero empirical depth with respect to both the classes. Clearly, the classifiers 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 empirical depth with respect to both of the competing classes. For classifying these observations, we rely on 1-nearest neighbor method. From the definition of SPD, however, it is quite transparent that if the distribution is not completely supported on a real line, SPD is always positive. As a result, both the observations 'A' and 'B' are correctly classified by SPD. This is a critical issue as both of HD and SD attain the value zero at points outside the support of the distribution when the support is bounded while SPD does not get affected by such problems in high dimensions.

4 Numerical results for equal prior cases

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 real data sets, we have reported the misclassification rates of some nonparametric methods to compare those with the performance of the maximum depth classifiers (d_D).

As we have already noted in Section 1, SPD is not invariant under general affine transformations. It will not be meaningful to compute SPD based on a multivariate data set when the co-ordinate variables are measured in different scales, and some standardization of the variables is necessary before computing SPD. One can use an estimate of the dispersion matrix based on empirical second moments to standardize the variables but we chose not to do that as our simulation studies involve distributions with heavy tails. We wanted our classifiers to be robust against the possible presence of outliers in the training data. On the other hand, use of more robust estimates of scatter matrix like the minimum volume ellipsoid or the minimum covariance determinant estimates (Rousseeuw, 1985; Rousseeuw and Van Drissen, 1999) will increase the computational cost substantially. In all our numerical studies, before computing SPD of an observation, we standardized the measurements variables in each class using marginal inter-quartile ranges. This enables us to use SPD even if different measurement variables are originally not in comparable scales. Throughout this section, prior probabilities for all competing populations are assumed to be equal.

4.1 Results on simulated data sets

As simulated examples, we consider some two-class problems, where both the populations are elliptically symmetric and they differ only in their location parameters $\boldsymbol{\mu}_1 = (0, \ldots, 0)$ and $\boldsymbol{\mu}_2 = (\mu, \ldots, \mu)$. The value of μ is taken to be 1 and 2 for our experiments. For proper evaluation of depth based classification methods, we investigate two extreme cases, where the observations are generated from multivariate normal distributions having exponential tails and multivariate Cauchy distributions having heavy polynomial tails. We consider $\mathbf{I}$= the identity matrix or $\boldsymbol{\Sigma}_0$ as the common scatter matrix of the two populations, where $\boldsymbol{\Sigma}_0$ is taken as $\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$ and $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 1 \end{bmatrix}$, respectively, for two and three dimensional problems. 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 4.1 and 4.2. For two dimensional problems, we report the performance of HD classifier based on its exact version, whereas the approximate version is used for $d = 3$. Due to computational difficulty, SD is used only in the case of bivariate problems.

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. When the population distributions are spherically symmetric (i.e. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$), SPD had a slight edge over the other two depth based classifiers, but when $\boldsymbol{\Sigma}_0$ is used as the common scatter matrix, especially in dimension 3, HD led to a slightly better performance. Recall that we used only the marginal quantiles to standardize the variables for SPD calculation, which do not take care of the correlations between the variables.

Standardization by any robust estimate of $\mathbf{\Sigma}^{-1/2}$ may lead to a better performance in such cases at the cost of increased complexity. However, the performance of LDA and QDA falls drastically when the observations are generated from a heavy tailed distribution like Cauchy. In this case, HD, SD and SPD clearly outperformed LDA and QDA. The error rates for depth based classifiers were reasonably close to the optimal Bayes risk while those for LDA and QDA were much higher for data simulated from Cauchy distributions.

Distribution	μ	$\mathbf{\Sigma}$	Bayes risk	n	LDA	QDA	SPD	HD	SD
Normal	1	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	30.86	100	31.02(0.13)	31.33(0.14)	31.61(0.14)	32.29(0.17)	32.51(0.19)
				200	31.11(0.16)	31.13(0.17)	31.68(0.17)	31.72(0.17)	31.88(0.17)
	2	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	15.86	100	16.07(0.11)	16.14(0.11)	16.92(0.11)	16.86(0.12)	16.88(0.12)
				200	15.99(0.11)	16.04(0.10)	16.89(0.12)	16.30(0.11)	16.40(0.12
Cauchy	1	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	35.24	100	45.68(0.67)	49.37(0.19)	38.36(0.25)	39.26(0.27)	39.31(0.28)
				200	44.78(0.73)	49.86(0.07)	36.68(0.21)	37.62(0.23)	37.70(0.24)
	2	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	25.01	100	40.83(1.19)	49.41(0.14)	27.78(0.23)	28.64(0.27)	28.80(0.28)
				200	38.65(1.18)	49.63(0.14)	26.79(0.19)	27.36(0.21)	27.51(0.19)

Table 4.1 : Misclassification rates (in %) for elliptic distributions with $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$ (dim. 2).

Distribution	μ	$\mathbf{\Sigma}$	Bayes risk	n	LDA	QDA	SPD	HD
Normal	1	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	21.45	100	21.87(0.16)	22.14(0.16)	25.99(0.17)	23.97(0.19)
				200	21.53(0.13)	21.70(0.13)	25.25(0.12)	22.77(0.14)
	2	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	5.70	100	5.94(0.08)	6.00(0.88)	9.98(0.12)	7.03(0.10)
				200	5.76(0.07)	5.80(0.08)	9.79(0.11)	6.36(0.09)
Cauchy	1	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	28.71	100	41.94(0.92)	49.46(0.22)	33.82(0.26)	34.01(0.32)
				200	42.59(0.86)	49.49(0.20)	33.08(0.18)	32.68(0.25)
	2	$\mathbf{I}$	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)
		$\mathbf{\Sigma}_0$	17.95	100	28.90(1.11)	48.32(0.52)	22.94(0.21)	22.81(0.28)
				200	30.69(0.98)	48.75(0.440	22.36(0.18)	21.72(0.25)

Table 4.2 : Misclassification rates (in %) for elliptic distributions with $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$ (dim. 3).

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 order of asymptotic accuracy of misclassification rates for suitable empirical depth based classifiers under appropriate regularity conditions.

Theorem 4.1 : *Suppose that the population density functions $f_1, f_2, \ldots, f_J$ satisfy the conditions of Theorem 2.1 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 $\Delta =$*

optimal Bayes risk. Then, in the equal prior cases, we have

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

for some appropriate function $\beta_{\mathbf{n}}$, which depends on the choice of the depth measure. 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, the error rate of the SPD classifier also satisfies the above inequality with $\beta_{\mathbf{n}}(t) = \prod_{j=1}^{J} max\{0, 1-2d\ e^{-n_j t^4/8d^2}\}$.

One should note that irrespective of the depth function, for every t, $\beta_{\mathbf{n}}(t)$ converges to 1 as $\min\{n_1, n_2, \ldots, n_J\}$ tends to infinity. This implies the asymptotic convergence of misclassification rates of maximum depth classifiers to the optimal Bayes risk.

4.2 Results from the analysis of benchmark data sets

We use three benchmark data sets, namely synthetic data, vowel recognition data and salmon data, for further illustration. Synthetic data and salmon data have the same number of observations for different populations, which justifies the use of equal priors for the competing classes. Since the sample sizes of the different classes in vowel data are not very different, we have taken the priors to be equal for evaluating the performance of the depth based classification techniques. Performance of LDA, QDA and nearest neighbor method (based on Euclidean distance and cross-validated choice of k) are also reported to facilitate the comparison. Since nearest neighbor and SPD classifiers are not affine invariant, we report the results for these methods both based on standardized and unstandardized version of the data sets. Usual sample dispersion matrix is used for this standardization. Unlike synthetic data and vowel data, salmon data does not have any 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 in Table 4.3. For synthetic data and vowel data, which have separate training and test sets, we report the test set misclassification errors for different classifiers. If a classifier leads to a test set error rate p, the corresponding standard error is taken as $\sqrt{p(1-p)/n}$, where n is the size of the test sample.

Synthetic data : This data set was used by Ripley (1994) and many other authors, who reported the error rates of different parametric and nonparametric classifiers. It is a well-known benchmark data set, where both the classes are equal mixtures of two bivariate normal distributions differing only in their location parameters. Parameters of these bivariate distributions were chosen to yield a Bayes risk of 8.0%. There is a training set consisting of 250 observations (125 from each class) and a test set consisting of 1000 cases (500 from each class). A scatter plot of this data set is given in Figure 4.1, where the dots ($\cdot$) and the crosses ($\times$) represent the observations from the two classes.

On this data set, LDA, QDA, classification tree and SPD classifier led to almost similar test set error rates. Misclassification rates for these methods were found to be 10.8%, 10.2%, 10.1% and 10.5%,

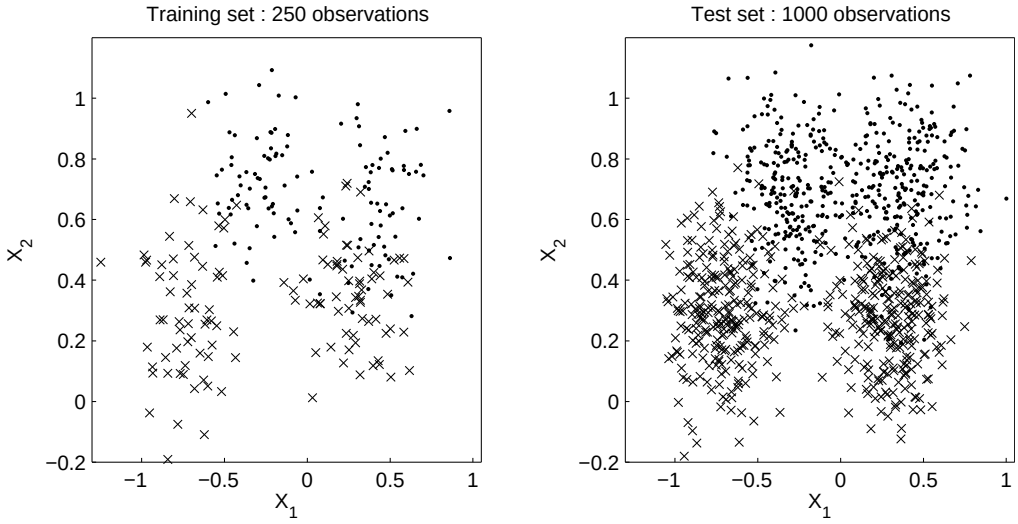

Figure 4.1 : Scatter plots for synthetic data

respectively. Nearest neighbor could achieve the best error rate of 8.7%. Misclassification rates for HD (error rate = 12.8%) and SD (error rate = 13.8%) were higher than that of the other classifiers. When the data were standardized using the sample dispersion matrix, SPD, classification tree and nearest neighbor method led to error rates of 10.5%, 12.0% and 11.7%, respectively.

Vowel recognition data : This data set 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 4.2 where the numbers represent the labels of different classes ('0' represents the 10-th class).

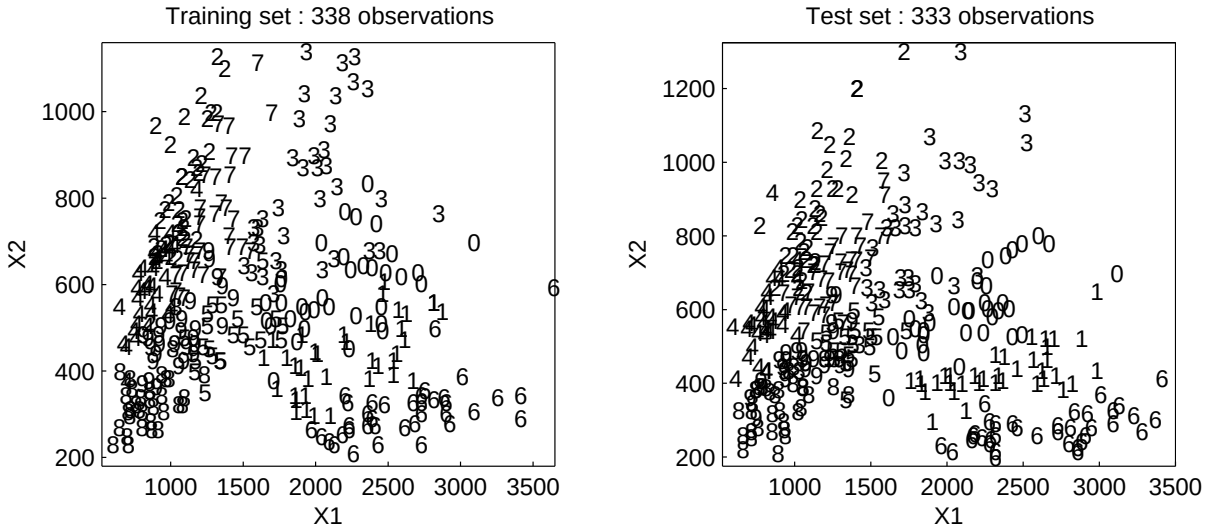

Figure 4.2 : Scatter plots for vowel data

Cooley and MacEachern (1998) used this data set to study the performance of different classification algorithms. On this data set, QDA could achieve an error rate =19.8%. Nearest neighbor method had a misclassification rate =21.9%. HD classifier and the classification tree method both led to the same error rate =23.7%. Misclassification rates for LDA and SD classifier were much higher than those of the other classifiers. SPD classifier led to an error rate =24.6%. However, when the data points were standardized using a pooled estimate of the dispersion matrix, this misclassification rate for SPD reduced to 21.3%. On this standardized data set, classification tree and nearest neighbor method had error rates 24.0% and 17.7%, respectively.

Salmon data : This data set is taken from Johnson and Wichern (1992, p. 520). It contains measurements on freshwater and marine water growth ring diameters on each of 100 salmon fish coming from Alaskan and Canadian water (50 from each population). A scatter plot of this data set is given in Figure 4.3.

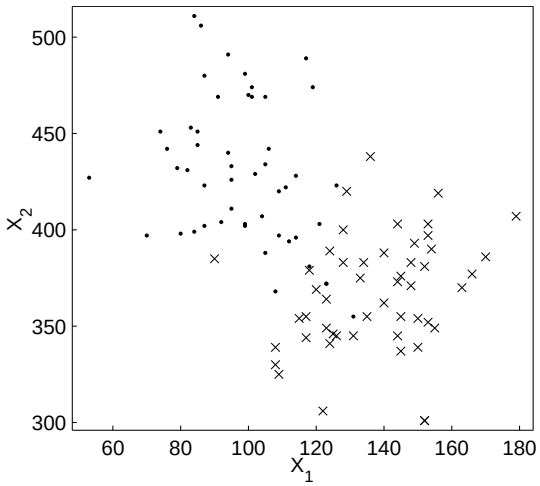

Figure 4.3 : Scatter plots for salmon data

On this data set, LDA (error rate =7.54%, S.E.=0.32%), QDA (error rate =7.23%, 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 had slightly higher error rates. For these two classifiers, the average misclassification rates were found to be 8.76% and 8.02%, respectively, with standard errors 0.41% and 0.36% in the respective cases. On the standardized version of the data, the nearest neighbor method and the SPD classifier could achieve average test set error rates 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.

Data sets	LDA	QDA	Nearest neighbor		SPD		HD	SD
			Original	Standardized	Original	Standardized		
Synthetic	10.8 (0.98)	10.2 (0.96)	8.7 (0.89)	11.7 (1.02)	10.5 (0.97)	10.5 (0.97)	12.8 (1.06)	13.8 (1.09)
Vowel	25.2 (2.38)	19.8 (2.18)	21.9 (2.27)	17.7 (2.09)	24.6 (2.36)	21.3 (2.24)	23.7 (2.33)	32.7 (2.57)
Salmon	7.54 (0.32)	7.23 (0.32)	8.02 (0.36)	8.11 (0.37)	7.46 (0.33)	7.42 (0.34)	7.32 (0.34)	8.76 (0.41)

Table 4.3 : Misclassification rates (in %) for benchmark data sets

From the analysis of these data sets, SPD and HD classifier seem to be better than SD classi-

fiers. Performance of these classifiers on the simulated and benchmark data sets was fairly competitive compared to the other parametric and nonparametric classification procedures. In most of the data sets, SPD led to smaller error rates than HD. In terms of computational cost, it had a clear edge over HD, SD and other maximum depth classifiers.

5 More on depth based nonparametric classification

In practice, different populations may have different priors, and they may not belong to the same family of elliptic distributions. In such cases, the maximum depth classifiers may not work well. In this section, we deal with such situations and propose another depth based classifiers which are capable to achieve reasonably lower misclassification rates under a more general set up that includes situations with unequal priors. These classifiers only assume the ellipticity of the data distribution to build up the decision rule and in that sense they are more flexible than maximum depth methods which need the populations to satisfy a location shift model to perform well.

Theorem 5.1 : *When the population distributions are elliptically symmetric, for any of MD, HD, SD, MJD, PD and SVD^{δ} with $\delta \geq 1$, 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*

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

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 the 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 based on population depth functions as we have already seen in preceding sections.

To construct a classification rule based on the training sample observations, one needs to find out appropriate sample analogs for $\theta_j\{D(j, \mathbf{x})\}$. Unfortunately, for most of the depth functions, $\theta_j\{D(j, \mathbf{x})\}$ is a complicated function of $D(j, \mathbf{x})$, and it is not easy 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 5.1 in the Appendix) in the case of elliptic distributions with location parameter $\boldsymbol{\mu}_j$ and scatter parameter $\boldsymbol{\Sigma}_j$, it is possible to have a simple 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 $\boldsymbol{\alpha}' \boldsymbol{\Sigma}_j^{-1/2}(\mathbf{X}_j - \boldsymbol{\mu}_j)$ for any α with $\|\alpha\| = 1$, and $\varrho_j(\cdot)$ being 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. One should notice that $\theta_j\{D(j, \mathbf{x})\}$ is nothing but the density function of the j^{th} population and our method tries to find out a consistent estimate of θ_j using depth. Such depth based

density estimation has also been investigated by Fraiman, Liu and Meloche (1997). Any moment based estimate of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ will make the estimate $\hat{\gamma}_j$ and the classifier sensitive to outliers and contaminated observations. Instead, it is better to plug in some robust estimates for these parameters. Here, we use half-space depth to estimate γ_j while kernel density estimation technique is used to estimate ϱ_j and to construct a new depth based nonparametric classifier (see the following section for detailed description). It requires a consistent estimate for $|\boldsymbol{\Sigma}_j|$ as well. Here, we bypass this estimation problem by writing the classifier in the form

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

where $C_1, C_2, \ldots, C_J$ are suitable constants. In practice, one can take $C_1 = 1$, and after finding some consistent estimates for γ_j and ϱ_j using the training sample, one will minimize the misclassification rate of the resulting classifier with respect to $C_2, C_3, \ldots, C_J$ to build up the final classification rule.

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 f_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 j^{th} population $(j = 1, 2)$. Next, we project the observations of f_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 $\gamma_j\{D(j, \mathbf{x}_0)\}\sqrt{\boldsymbol{\alpha}' \boldsymbol{\Sigma}_j \boldsymbol{\alpha}}$. The following theorem establishes the consistency of this estimate under appropriate condition.

Theorem 5.2 : *Suppose that* $\mathbf{X}$ *has an elliptically symmetric density with location parameter* $\boldsymbol{\mu}$ *and scatter matrix* $\boldsymbol{\Sigma}$. *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 as that of* $\mathbf{X}$. *Also, define* $\xi_{p,\boldsymbol{\alpha},n}$ *as the* p*-th* $(0 < p < 1)$ *empirical quantile of* $\boldsymbol{\alpha}'\mathbf{X}$ *for some* $\boldsymbol{\alpha}$ *with* $\|\boldsymbol{\alpha}\| = 1$. *Then, as* $n \to \infty$, $(\xi_{1-\delta_n,\boldsymbol{\alpha},n} - \xi_{\delta_n,\boldsymbol{\alpha},n})/2 \xrightarrow{a.s.} \{(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}^{1/2} \sqrt{\boldsymbol{\alpha}' \boldsymbol{\Sigma} \boldsymbol{\alpha}}$.

This estimation procedure can be repeated using a number of different directions $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, ..., $\boldsymbol{\alpha}_S$ and the average of these estimates can be 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 $\lambda_j = \sum_{s=1}^{S} \sqrt{\boldsymbol{\alpha}_s' \boldsymbol{\Sigma}_j \boldsymbol{\alpha}_s}/S$. Our empirical study suggest that the final classifier is not much sensitive to the choice of $\boldsymbol{\alpha}_s$'s. For all data analytic purposes in this article, we have used $S = d$, where $\boldsymbol{\alpha}_s$'s are taken as the unit vectors along the co-ordinate axes. It should also be noted that the form of this new depth based 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 $\hat{\mathbf{v}}_{21}, \hat{\mathbf{v}}_{22}, \ldots, \hat{\mathbf{v}}_{2n_2}$ are obtained, where $\hat{\mathbf{v}}_{ji} = (\hat{v}_{ji}^{(1)}, \hat{v}_{ji}^{(2)})$ denotes the estimate of $(v_{ji}^{(1)}, v_{ji}^{(2)})$, the re-scaled Mahalanobis distances of $\mathbf{x}_{ji}$ from 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 can estimate the density function (Υ_j, say) of the re-scaled

Mahalanobis distance by any appropriate nonparametric methods [note that $\Upsilon_j(v) = \lambda_j \ \varrho_j(\lambda_j \ v)$]. If kernel methods are used for density estimation, one has to find out appropriate bandwidth parameters as well. Instead of using bandwidths that minimizes 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 Ghosh and Chaudhuri, 2004a). However, to find out these 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 of 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 5.2, 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 5.1 in the Appendix) $\hat{\Upsilon}^*_{jh_{jn_j}}(\hat{v}_0^{(j)})$ 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 a 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 the new depth based classifier of the form

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

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

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 more general models, where the prior probabilities may or may not be equal. Density estimation using depth (see also Fraiman, Liu and Meloche, 1997) requires only the one dimensional densities to be estimated, and just like the projection pursuit method it helps to avoid the problem of sparsity in higher dimensions. Further, it makes the convergence of the density estimates faster than that in d-dimensional kernel density estimation. However, it is computationally difficult to minimize the error rate simultaneously with respect to $C_1, C_2, \ldots, C_J$ as well as bandwidth parameters $h_1, h_2, \ldots, h_J$. 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 method of majority voting (see e.g., Friedman, 1996) to arrive at the final classification.

6 Numerical results

In this section, we use some simulated and benchmark data sets for further illustration. Along with the results of the proposed new classifier (henceforth called HD^*) based on half-space depth, 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 real data set to facilitate comparison.

6.1 Simulated examples on unequal prior cases

For simulation experiments, we consider two dimensional problems only. We choose the same two-dimensional examples with normal and Cauchy populations as discussed in Section 4.1, where the two competing populations satisfy a location shift model. But this time, we use unequal priors π_1 and $1 - \pi_1$ for the two competing populations. For each experiment, the reported results (see Table 6.1) 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 QDA and HD^* were also fairly comparable. In the case of Cauchy distribution, the classifier HD^* outperformed LDA and QDA like the maximum depth classifiers considered earlier.

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)	30.72(0.28)
				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 6.1 : Misclassification rates (in %) on elliptic distributions when $\pi_1 \neq \pi_2$ (dimension = 2)

6.2 Examples with equal priors but different scatters and shapes

As we have already pointed out, unlike the maximum depth classifiers, HD^* can work well even when the competing populations have different shapes and scatter matrices. Here, we consider some example of that kind to illustrate the utility of the classifier HD^*.

We begin with some examples, where the two elliptic populations differ only in their location and scatter parameters. Let us consider the two-dimensional examples with normal and cauchy distributions as discussed in Section 3. The location parameters of these distributions are chosen as before with $\mu = 1$ and 2, but this time the we take different scatter matrices ($\mathbf{\Sigma}_1 = \mathbf{I}$ and $\mathbf{\Sigma}_2 = 4\mathbf{I}$) for the two populations. Each experiment is carried out 100 times as before and the results for different classification methods are reported in Table 6.2. Since the optimal class boundaries are quadratic, as it is expected, QDA outperformed the other classifiers when the underlying distributions are normal. The classifier HD^*, in these cases, performed significantly better than the maximum depth methods, and it could nearly match the performance of QDA when we have relatively larger training samples. Once again, for cauchy distribution, depth based classification procedures outperformed the traditional approaches.

Distribution	μ	Bayes risk	n	LDA	QDA	SPD	HD	SD	HD*
Normal	1	22.03	100	30.53(0.15)	22.40(0.13)	37.03(0.32)	36.85(0.32)	37.20(0.31)	28.24(0.36)
			200	30.21(0.15)	22.23(0.13)	36.29(0.26)	36.18(0.25)	36.38(0.26)	25.81(0.27)
	2	13.31	100	16.24(0.11)	13.56(0.11)	19.85(0.26)	19.71(0.26)	20.27(0.26)	16.47(0.24)
			200	16.11(0.10)	13.47(0.10)	19.51(0.18)	19.39(0.19)	19.48(0.18)	15.22(0.20)
Cauchy	1	30.92	100	44.93(0.76)	47.64(0.32)	40.50(0.35)	40.62(0.37)	41.03(0.39)	36.07(0.31)
			200	45.39(0.76)	48.27(0.21)	40.27(0.23)	40.34(0.25)	40.50(0.26)	33.98(0.22)
	2	22.27	100	38.40(0.11)	46.94(0.37)	29.98(0.38)	30.39(0.40)	30.62(0.38)	27.82(0.27)
			200	39.66(0.11)	48.61(0.23)	29.82(0.29)	30.18(0.30)	30.37(0.32)	25.83(0.17)

Table 6.2 : Misclassification rates (in %) for elliptic distributions with $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 4\mathbf{I}$.

Advantage of the classifier HD^* becomes more evident when the competing populations are of very different shapes. Here, we consider an example with two classes, where both populations are distributed as $N_2(0, 0, 1, 1, 0)$ but the second one is truncated at $x_1^2 + x_2^2 > 4$. Though both the populations are elliptically symmetric, they have very different structure. Clearly, the optimal class boundary for this problem is $x_1^2 + x_2^2 = 4$. To study the performance of different depth based methods, we generate equal number of observations from the two classes to form a training set and a test set each of size 1000. We generate 100 such samples and use different classifiers to classify them. Average test set error rates over these 100 trials and their corresponding standard errors are computed for different methods. In this example, QDA led to an average test set error rate of 9.87% but LDA and the three maximum depth classifiers misclassified nearly 50% of the test set observations. The classifier HD^* performed better than these classifiers. It could lead to an average test set error rate of 8.31% with a standard error of 0.06%. When the normal populations are replaced by Cauchy distributions, along with LDA and maximum depth classifiers, QDA also failed to capture the optimum class boundary. But, even in that case, the classifier HD^* performed well and it could nearly achieve the optimal Bayes risk.

	Bayes risk	LDA	QDA	SPD	HD	SD	HD*
Truncated Normal	6.77	49.59(0.16)	9.87(0.09)	49.82(0.03)	50.06(0.04)	50.07(0.04)	8.31(0.06)
Truncated Cauchy	22.36	46.31(0.63)	48.46(0.14)	49.83(0.03)	49.98(0.04)	50.10(0.04)	25.08(0.23)

Table 6.3 : Misclassification rates (in %) when competing populations have different shapes.

6.3 Results on biomedical data

We now consider the "biomedical data set" (see Cox, Johnson and Kafadar, 1982) to compare the performance of different classification methods. This data set (available at CMU data sets archive) contains information on four different measurements on each of the 209 blood samples (134 for "normals" and 75 for "carriers"). Out of these 209 observations, 15 have missing values. We have removed those observations and applied the classification methods on the remaining 194 cases (127 for "normals" and 67 for "carriers"). Biomedical data does not have separate training and test sets, and we formed these sets by random partitioning of the data. This partitioning 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. Since the priors are quite different, maximum depth classifiers were not used for classification.

In this data set, LDA and QDA led to error rates of 15.75% (S.E.=0.32%) and 12.26% (S.E.=0.21%) respectively. HD^* classifier could achieve an error rate 13.98% with a standard error 0.31%. The usual k-nearest neighbor classifier based on Euclidean distance and cross-validated estimate of k had an error rate 13.22% (S.E.=0.27%). Here also we applied the nearest neighbor method after standardizing the observations by pooled dispersion matrix. On this standardized data set, the nearest neighbor method led to a worse misclassification rate =16.26% with a standard error =0.30%.

7 Concluding remarks

Among the maximum depth based classifiers discussed in this article, SPD seems to have some definite advantages. Not only it requires less computation but also gets less affected by the "problem of zero depth" pointed out in Section 3 for high dimensional data. In most of the examples in Section 4, SPD led to better performance than the other maximum depth classifiers in equal prior problems. Further, compared to the performance of other well-known nonparametric methods like nearest neighbors and classification trees, maximum depth classifiers, especially the SPD classifier, led to fairly satisfactory performance in all the data sets that we have analyzed. From the numerical results, it seems that SPD classifier has the potential to be used as a computationally efficient robust alternative to many traditional methods of discriminant analysis. However, when the distributions of data in competing populations have very different shapes, all of the maximum depth classifiers may have very poor performance.

This article also throws some light on possible generalization of depth based classifiers for unequal prior cases. The classifier HD^* based on half-space depth is computationally more expensive than maximum depth classifiers. But HD^* is more flexible in nature and requires less assumption on the data distributions.

Before we finish, we would like to address the issue of computational cost for the classifier HD^*. It involves two major computational steps : (i) depth computation and estimation of re-scaled Mahalanobis distance for all observations (ii) kernel density estimation and simultaneous estimation of optimum bandwidth pair (h_1, h_2) and the corresponding cut off value C^* by leave-one-out cross-validation. Consider now the training sample sizes $n_1, n_2, \ldots, n_J$ such that $n = n_1 + \ldots, n_J$, and we assume n_i/n remains bounded away from zero as n tends to infinity for all $1 \leq i \leq J$. On a d-dimensional data set, use of available algorithms for exact depth depth computation require $O(n^d logn)$ calculations (see e.g., Rousseeuw and Ruts, 1996; Rousseeuw and Struyf, 1998) to complete the first step mentioned above. For $d > 2$, we have used an approximate algorithm for depth computation, which has been described in detail in Ghosh and Chaudhuri (2004), and this reduces the computational cost significantly. It is an iterative algorithm, and for each observation it requires $O(n)$ calculations to perform one iteration. For the second step mentioned above, one has the evaluate the misclassification probability for varying choices of bandwidth pairs (h_1, h_2) by leave-one-out method. For a given value (h_1, h_2), it requires $O(n^2)$ computations in a two-class problem to calculate the misclassification probability in a leave-one-out method. Then, one has to repeat this computation of misclassification probability over different

choices of bandwidth parameters (varying over a grid of values of h_1 and h_2) to find out the optimal bandwidth parameters and the cut-off value C^*. For a J class problems, this has to be carried out $J(J-1)/2$ times taking all possible pairs of classes. For the biomedical data set with two classes, a Pentium-4 processor took less than four minutes to complete all those above mentioned operations when a search over 100×100 values of (h_1, h_2) was used for finding the best pair of bandwidths.

Acknowledgement : Authors are thankful to two anonymous referees, who carefully read an earlier version of the paper and made several helpful comments.

Appendix : Proofs

Proposition 2.1 : 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 2.1 : 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 these points 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 of $\sum_{i=1}^{4} \frac{\mathbf{x}^{(i)} - \mathbf{x}_1}{\|\mathbf{x}^{(i)} - \mathbf{x}_1\|} f(\mathbf{x}^{(i)})$ 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, integrating 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. □

Some general observations regarding Theorems 2.1-2.3 : In order to prove Theorems 2.1-2.3, 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 $\mathbf{x}$ with respect to 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 almost sure pointwise convergence of empirical depth functions to population depth functions, the result will follow as an immediate consequence of dominated convergence theorem provided that 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 whenever the populations are elliptic differing only in their location parameters. 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 2.1 : Results on uniform convergence of the empirical versions of 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), and the theorem follows from that. □

Proof of Theorem 2.2 : Uniform convergence of the empirical version of spatial depth to its population analogue follows from the work of Kolchinskii (1997) and Serfling (2002), and that proves the theorem. □

Proof of Theorem 2.3 : Since the populations satisfy the location shift model, it is not necessary to have the term $|\mathbf{\Sigma}|^{1/2}$ 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 $\mathbf{x}$, the empirical version of SVD^{δ} converges almost surely to its population counter part. This completes the proof. □

Proof of Theorem 4.1 : Note that under the given conditions the population depth based classifiers turn out to be the optimal Bayes classifier, and Δ can be expressed as

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

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

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

Here, the inside probability (P) denotes the probability with respect to the distribution of $\mathbf{X}_j \sim f_j$, the outer probability (P) is with respect to the distribution of all the $\mathbf{x}_{ji}$'s, and $\mathbf{x}$ is fixed. Now, the set of hyperplanes $\{\mathbf{X} : l^{'}(\mathbf{X}-\mathbf{x}) = 0\}$ in R^d with varying l has VC dimension d (see e.g., Pollard, 1984; Vapnik, 1998). 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 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 $\epsilon = D^{01}(\mathbf{x})/2$.

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

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

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

(*ii*) (**The case of** SD) : The sample version of simplicial depth is a U-statistic with a bounded kernel function. Therefore, using Hoeffding's inequality (see e.g., Hoeffding, 1963; Serfling, 1980) we have

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

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

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

(*iii*) (**The case of** SPD) : For the ease of 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 $\boldsymbol{\mu}_{\mathbf{Z}} = E(\mathbf{Z})$. Since $\|\bar{\mathbf{z}}_{n_j}\|$ and $\|\boldsymbol{\mu}_{\mathbf{Z}}\|$ both are positive, we have

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

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

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

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

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

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

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

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

□

Proof of Theorem 5.1 : Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ be the location parameter and the scatter matrix of the j^{th} population which has a density function f_j. Define $R_j = \left\{(\mathbf{X}_j - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{X}_j - \boldsymbol{\mu}_j)\right\}^{1/2}$, where $\mathbf{X}_j \sim f_j$.

When, f_j is elliptically symmetric, the distributions of R_j is given by (see e.g., Fang, Kotz and Ng, 1989)

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

where $r_j = \{(\mathbf{x}-\boldsymbol{\mu}_j)^{'}\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\}^{1/2}$, the Mahalanobis distance of $\mathbf{x}$ from $\boldsymbol{\mu}_j$. Clearly, $\pi_j f_j(\mathbf{x}) > \pi_i f_i(\mathbf{x}) \Leftrightarrow \pi_j|\mathbf{\Sigma}_j|^{-1/2}\ \varrho_j(r_j)/r_j^{d-1} > \pi_i|\mathbf{\Sigma}_i|^{-1/2}\ \varrho_i(r_i)/r_i^{d-1}$, 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 as

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

□

Lemma 5.1 : Suppose that $\mathbf{X}$ follows an elliptic distribution G, and $\mathbf{X} = \boldsymbol{\mu} + \mathbf{\Sigma}^{1/2}\mathbf{U}$, where $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are some location and scatter parameters, and $\mathbf{U}$ is a spherically distributed random vector. Then, for any given $\mathbf{x}$ and any given direction $\boldsymbol{\alpha}$ ($\|\boldsymbol{\alpha}\| = 1$), we have

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

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

Proof of Lemma 5.1 : Half-space depth of $\mathbf{x}$ with respect to G can be expressed as

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

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

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

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

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

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

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

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

Lemma 5.2 : 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 5.2 : 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 4.1), 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 5.2 : From Lemma 5.2, it is easy to see that $\xi_{\delta_n,\boldsymbol{\alpha},n} \xrightarrow{a.s} \xi_{\delta,\boldsymbol{\alpha}}$ and $\xi_{1-\delta_n,\boldsymbol{\alpha},n} \xrightarrow{a.s} \xi_{1-\delta,\boldsymbol{\alpha}}$, where $\xi_{p,\boldsymbol{\alpha}}$ is the p-th quantile of $\boldsymbol{\alpha}'\mathbf{X}$. Now, from Lemma 5.1, it follows that

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

Proposition 5.1 : Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are observations from G (as described in Lemma 5.1), and $\xi_{p,\boldsymbol{\alpha}}$ and $\xi_{p,\boldsymbol{\alpha},n}$ have the same meaning as in Theorem 5.2. For some given $\boldsymbol{\alpha}$ $(\|\boldsymbol{\alpha}\| = 1)$, define $v_i = \{(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\}^{1/2} \sqrt{\boldsymbol{\alpha}' \boldsymbol{\Sigma} \boldsymbol{\alpha}}$ and $\hat{v}_i = (\xi_{1-\delta_{i_n},\boldsymbol{\alpha},n} - \xi_{\delta_{i_n},\boldsymbol{\alpha},n})/2$ 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 v_i's have the 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 it satisfies $\int |t|^3 K^2(t) dt < \infty$.

(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 5.1 : 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 5.1, it is easy to verify that $|v_0 - \hat{v}_0| \leq \frac{1}{2}\{|\xi_{\delta,\boldsymbol{\alpha}} - \xi_{\delta_n,\boldsymbol{\alpha},n}| + |\xi_{1-\delta,\boldsymbol{\alpha}} - \xi_{1-\delta_n,\boldsymbol{\alpha},n}|$. Therefore,

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

for some $\eta > 0$ and $m > 0$ as chosen in Lemma 5.2. This implies that

$$P\{|\hat{\Upsilon}_{h_n}(v_0) - \hat{\Upsilon}^*_{h_n}(\hat{v}_0)| > \epsilon\} < 8(n^{d+1} e^{-nh_n^2\eta^2/2} + e^{-nh_n^2 m^2/2}),$$

and hence $|\hat{\Upsilon}^*_{h_n}(\hat{v}_0) - \hat{\Upsilon}_{h_n}(v_0)| \xrightarrow{P} 0$ as $n \to \infty$. Now, under the assumed conditions, the expectation and the variance of $\hat{\Upsilon}_{h_n}(v_0)$ are of the form (see e.g., Ghosh and Chaudhuri, 2004)

$$E\{\hat{\Upsilon}_{h_n}(v_0)\} = \Upsilon(v_0) + O(h_n^2) \quad \text{and} \quad Var\{\hat{\Upsilon}_{h_n}(v_0)\} = O(n^{-1}h_n^{-1}),$$

which implies that $|\hat{\Upsilon}_{h_n}(v_0) - \Upsilon(v_0)| \xrightarrow{P} 0$ as $n \to \infty$. Therefore, $\hat{\Upsilon}^*_{h_n}(\hat{v}_0)$ converges to $\Upsilon(v_0)$ in probability. □

References

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

[2] Chaudhuri, P. (1996) On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, **91**, 862-872.

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

[4] 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.

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

[6] Cox, L. H., Johnson, M. M. and Kafadar, K. (1982) Exposition of Statistical Graphics Technology, *ASA Proc. Statist. Comp. Section*, 55-56.

[7] Croux, C. and Dehon, C. (2001) Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics*, **29**, 473-492.

[8] Donoho, D. (1982) Breakdown properties of multivariate location estimators. *Ph.D. qualifying paper, Dept. of Stat., Havard University.*

[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. Wiley, New York.

[10] Fang, K-T., Kotz, S. and Ng, K. W. (1989) *Symmetric multivariate and related distributions.* Chapman and Hall, London.

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

[12] Fraiman, R., Liu, R.Y. and Mechole, J. (1997) Multivariate density estimation by probing depth. L_1 *Statistical Procedures and Related Topics. IMS Lecture Notes* (Y.Dodge ed.), **31**, 415-430.

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

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

[15] Ghosh, A. K. and Chaudhuri, P. (2004b) On data depth and distribution free discriminant analysis using separating surfaces. To appear in *Bernoulli*.

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

[17] He, X. and Fung, W. K. (2000) High breakdown estimation or multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, **72**, 151-162.

[18] Hoberg, R. (2000) Cluster ananlysis based on data depth. *Data Analysis, Classification and Related Methods* (H.A.L. Kiers, J.P. Rasson, P.J.F. Groenen and M. Schader eds.), Springer, Berlin, 17-22.

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

[20] Hubert, M. and Van Driessen, K. (2003) Fast and robust discriminant analysis. To appear in *Computational Statistics and Data Analysis*.

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

[22] Jornsten, R. (2004) Clustering and classification based on the L1 data depth. *Journal of Multivariate Analysis*, **90**, 67-89.

[23] Koltchinskii, V. I. (1997) M-estimation, convexity and quantiles. *Ann. Statist.*, **25**, 435-477.

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

[25] Liu, R. and Singh, K. (1993) A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.*, **88**, 252-260.

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

[28] Mizera, I. (2002) On depth and deep points : a calculus. *Ann. Statist.*, **30**, 1681-1736.

[29] Mosler, K. (2002) *Multivariate dispersions, central regions and depth*. Springer Verlag, New York.

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

[31] Oja, H. (1983) Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.*, **1**, 327-332.

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

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

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

[35] Rousseeuw, P. (1985), Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications* (W. Grossmann, G. Pflug, I. Vincze and W. Wertz eds.), Reidel, Dordrecht, pp. 283-297.

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

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

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

[39] Rousseeuw, P.J. and Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212-223.

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

[41] Serfling, R. (2002) A depth function and a scale curve based on spatial quantiles. *In Statistics and Data Analysis based on L_1-Norm and Related Methods* (Y. Dodge ed.), Birkhaeuser, 25-38.

[42] Singh, K. (1991) A notion of majority depth. *Tech. Report, Dept. of Statistics, Rutgers University.*

[43] Stahel, W. A. (1981) Breakdown of covariance estimators. *Research Report 31, Fachgruppefur Statistik, ETH, Zurich.*

[44] Tukey, J. (1975) Mathematics and the picturing of data. *Proceedings of the 1975 International Cong. of Math., Vancouver*, 523-531.

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

[46] 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.

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

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