On robust classification using projection depth

Subhajit Dutta and Anil K. Ghosh

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute

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

Email : subhajit_r@isical.ac.in, akghosh@isical.ac.in

Abstract

Data depth is a powerful tool for robust multivariate analysis with wide spread applications in different areas of statistics. Various measures of depth have been used for supervised classification as well. This paper uses projection depth for this purpose and investigates the asymptotic optimality of the resulting classifier under appropriate regularity conditions. Some simulated examples and benchmark data sets are analyzed to show the utility of the proposed method.

Index terms : Bayes risk, bandwidth, cross-validation, elliptic symmetry, kernel density estimation, misclassification rate, multi-scale smoothing, projection depth.

1 Introduction

Data depth measures the centrality of an observation w.r.t. a multivariate distribution or w.r.t. a multivariate data cloud. Over the last two decades, it has emerged as a powerful tool for generalizing different robust univariate statistical methods into multivariate set up. For $d > 1$, the lack of natural ordering in $\mathbb{R}^d$ makes it difficult to adopt any robust rank based statistical method for multivariate data analysis. The concept of data depth helps to

overcome this limitation by providing a center outward ordering of multivariate observations. Therefore, statistical depth functions have found increasingly many users in the area of non-parametric analysis. For instance, they have been used to compute multivariate location and scatter [27, 40], for the test of elliptic symmetry [20, 3], detection of outliers [24] and other problems of statistical inference [26, 32].

Since data depth provides a center outward ordering of multivariate data, it was naturally adopted for supervised and unsupervised classification as well. Jornsten [24] used L_1 depth (L_1D) [45, 40] for classification and clustering of microarray gene expression data. Zonoid depth [30] was used in [18, 19] for robust clustering and classification. Christmann *et. al.* [4] used regression depth (RD) [36] for linear and nonlinear classification. Along with RD, Ghosh and Chaudhuri [12] used half-space depth (HD) [43] for robust linear and quadratic classification. They have also developed some depth based nonparametric classifiers known as the maximum depth classifiers [13], where they used Mahalanobis depth (MD) [29], HD, simplicial depth (SD) [25] and L_1D for classification. Recently, band depth was introduced in [28] for classification of functional data.

In this article, first we will use Projection Depth (PD) [48] to develop a maximum depth classifier, and then we will modify it to construct a generalized depth based classifier. Recall that a maximum depth classifier classifies an observation $\mathbf{x}$ to the class in which it has the maximum depth. We can use PD for this purpose. To define PD of $\mathbf{x}$ w.r.t. a multivariate distribution F, one needs to measure the outlyingness of $\mathbf{x}$ w.r.t. F. It is given by

$$\mathrm{O}(\mathbf{x}, F) = \sup_{\alpha:\|\alpha\|=1} \{|\boldsymbol{\alpha}^{'}\mathbf{x} - m_F(\boldsymbol{\alpha}^{'}\mathbf{X})|/\sigma_F(\boldsymbol{\alpha}^{'}\mathbf{X})\},$$

where m_F and σ_F stand for some univariate measures of location and scatter, respectively.

Zuo and Serfling [48] used median and MAD (median absolute deviation) as these measures. However, one can use other measures as well. Projection depth of the observation, $\mathrm{PD}(\mathbf{x}, F)$, is obtained from $\mathrm{O}(\mathbf{x}, F)$ using the formula $\mathrm{PD}(\mathbf{x}, F) = 1/\{1 + \mathrm{O}(\mathbf{x}, F)\}$. Empirical version of this outlyingness and depth can be obtained if we replace F by its empirical analog F_n, which puts a mass of $1/n$ on each of the n data points. Like some other depth functions, PD possesses some nice properties. In fact, it satisfies all the four desirable properties of depth functions mentioned in [48], i.e., affine invariance, maximality at center, monotonicity relative to deepest point and vanishing at infinity. In [17, 49], the authors also showed the convergence of the empirical PD contours to their population analogs under some regularity conditions. Zuo [50], under some mild conditions, proved the uniform continuity of PD.

If an observation is a proper representative of a class, it is expected to have higher depth w.r.t. that class. On the contrary, if it is not from that class, it is expected to be an outlier for that class, and it will have depth close to zero w.r.t. that class. Therefore, it is somewhat reasonable to classify an observation to the class for which it has the maximum depth. In principle, almost all depth functions can be used for this maximum depth classification, but because of computational difficulty, it is not feasible to use some of the depth functions like SD [25], simplicial volume depth [31], zonoid depth [30] and majority depth (MJD) [42] in high dimensions. MD is the easiest one to compute, but this computational simplicity arises because of the use of moment based estimates of the mean vectors and the dispersion matrices, which are not robust. To make it robust, one needs to compute robust estimates for the location and the scatter, but at the cost of additional computations. For instance, the minimum-covariance determinant (MCD) estimators [35, 23] can

be used for this purpose. One can also use other robust and consistent estimates as well [44, 6]. L_1 depth is also easy to compute, but unlike other depth functions, its usual version does not have the affine invariance property. To find an affine invariant robust version $L_1D(\mathbf{x}, F) = 1 - \|E_F\{\mathbf{\Sigma}_F^{-1/2}(\mathbf{x} - \mathbf{X})/\|\mathbf{\Sigma}_F^{-1/2}(\mathbf{x} - \mathbf{X})\|\}\|$, here also, one needs to plug-in a robust estimate for the scatter matrix $\mathbf{\Sigma}_F$. HD satisfies all four desirable properties of a depth function, and it can also be computed in high dimensions [37, 2, 12]. But, one major limitation of HD is the stepwise nature of its empirical version. Here, an observation can have the maximum depth w.r.t. more than one classes. This problem becomes more serious when the observation lies on or outside the convex hulls formed by the training data from different classes. In such cases, it has zero depth w.r.t all competing classes. Maximum depth classifier based on HD often fails to correctly classify these observations. We also have similar problems for SD and MJD. Unlike these depth functions, both the empirical and the population versions of PD are continuous in $\mathbf{x}$, and they are always positive. So, one does not have to deal with ties and observations with zero depth.

The organization of the paper is as follows. In the next section, we use PD as a tool for robust maximum depth classification and study the theoretical properties of the resulting classifier (called the PD classifier). Some simulated data sets are also analyzed in this section to show the utility of this classifier. If the population distributions are elliptically symmetric, and if they satisfy a location shift model, this classifier works well, and it is capable of achieving error rates close to the Bayes risk when the prior probabilities of different classes are equal. In Section 3, we propose a generalization of this PD classifier to take care of a more general set up. To develop this generalized PD classifier, we need to estimate the density function

of the underlying distributions using the density of PD. Usual kernel density estimation [41] is used for this purpose, where the bandwidth h is chosen by cross-validation technique [34]. Asymptotic optimality of the error rate of this generalized PD classifier has also been derived here. Section 4 deals with multi-scale analysis, where instead of choosing a single h, we consider the results for different values h and aggregate them judiciously to yield better performance. Section 5 analyzes some benchmark datasets to show the usefulness of our proposed classifiers. Finally, Section 6 contains a brief summary of the entire work. All proofs and mathematical details are given in the Appendix.

2 Projection depth classifier

If we have $n_1, n_2, \ldots, n_J$ observations from J competing classes, the PD classifier is given by

$$d_{1n}(\mathbf{x}) = \arg\max_j \mathrm{PD}(\mathbf{x}, F_{jn_j}) = \arg\min_j \mathrm{O}(\mathbf{x}, F_{jn_j}),$$

where F_{jn_j} is empirical version of the j-th population distribution F_j $(j = 1, 2, \ldots, J)$ and $n = (n_1, n_2, \ldots, n_J)$. Note that if the population distributions are elliptically symmetric, and they follow a location shift model, for any observation $\mathbf{x}$, the ordering of $\mathrm{O}(\mathbf{x}, F_j)$ for different classes is the same as that of the corresponding Mahalanobis distances. One can also notice that under this set up, the Mahalanobis distance classifier is the Bayes classifier when the priors are equal. So, in this case, the error rate of the PD classifier converges to the Bayes risk as the training sample size increases. This result is presented in Theorem 1.

Theorem 1 *: If the population distributions are elliptically symmetric and they satisfy a location shift model (i.e.* $f_j(\mathbf{x}-\boldsymbol{\mu}_j) = f(\mathbf{x})$ *for some common density function* f *and location*

parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_J$*), the misclassification rate of the PD classifier* $d_{1n}(\cdot)$ *converges to the Bayes risk as* $\min\{n_1, n_2, \ldots, n_J\} \to \infty$.

2.1 Performance on simulated data sets

In this section, we analyze some simulated data sets to study the performance of the PD classifier. In order to compare it with other maximum depth classifiers, here we restrict ourselves to two class problems in two dimensions so that computationally expensive depth functions like SD can also be used for comparison. Nedler-Mead algorithm [46] available in R-package can be used for computation of PD. However, for two dimensional problems, one can search the direction $\boldsymbol{\alpha} = (sin\theta, cos\theta)$ over finer grids to compute the empirical PD of an observation. Since it requires less computation, in this article we have adopted this method for PD calculation. We used both MAD and quartile deviation (QD) as the measure of scale, but there was no visible difference in the performance of the resulting classifier. Therefore in this section, for PD classification, we have reported the result based on QD only. Note that if we use MD for maximum depth classification, it leads to usual linear discriminant analysis (LDA). Here we have reported the error rates for LDA and also for its robust versions, where MCD estimates of the location and the scatter are plugged in. As it has been mentioned before, any measure of univariate location and scale can be used to define PD. Here, we used median as the measure of location and two different MCD estimates for scatter, one of which is computed based on 50% observations (which has the highest breakdown) and the other one is based on 75% observations (as suggested in [23] for good finite sample efficiency). We will refer to these two classifiers as $\text{MD}_{1/2}$ classifier and $\text{MD}_{3/4}$ classifier, respectively.

Though these estimates converge to scalar multiples of the corresponding scatter matrices, under the assumptions of Theorem 1, these constants are same for all populations, and they do not affect the performance of the maximum depth classifier. Note that for affine invariant version of L_1D, one needs to estimate the scatter matrix as well, and here we have used the MCD estimate with highest breakdown for this purpose. As expected, throughout this section, prior probabilities of all competing classes are taken to be equal.

We begin with an example with two normal distributions having the same dispersion matrix $\mathbf{I_2}$, the two dimensional identity matrix, but different location parameters $(0, 0)$ and (μ, μ). We carried out our experiments with two different choices of μ (1 and 2). Each time, we generated a training set of size 100 (or 200) and a test set of size 500 taking equal number of observations from the competing classes. This procedure is repeated 500 times, and the average test set error rates of different depth based classifiers and their corresponding standard errors are reported in Table 1. Bayes error rates are also reported to facilitate the comparison. As we have already mentioned, for HD and SD, we may have some observations, which have zero depth w.r.t. all competing classes. Those observations were classified using the nearest neighbor [5] algorithm. Otherwise, the performance of these two depth based methods would have been worse. As expected, LDA led to the best error rate in all these examples, and these error rates were very close to the corresponding Bayes risks. Performance of its robust versions ($MD_{1/2}$ and $MD_{3/4}$), and that of L_1D and PD was also fairly competitive. We repeated the same experiment using Cauchy instead of normal distributions. In the presence of extreme valued observations from heavy-tailed distributions, LDA failed to perform satisfactorily. In all these cases, the performance of all other depth based classifiers

was significantly better than LDA, which shows the robustness of these depth based methods. We carried out our experiments with different choices of the location and the scatter parameters, but the basic finding remained the same.

Table 1 : Error rates (in %) of different maximum depth classifiers and their standard errors.

		Normal		Cauchy	
		$\mu = 1$	$\mu = 2$	$\mu = 1$	$\mu = 2$
Bayes risk $\longrightarrow$		23.98	7.87	30.40	19.58
LDA	$n_1 = n_2 = 50$	24.18 (0.09)	8.05 (0.06)	42.61 (0.43)	32.27 (0.60)
	$n_1 = n_2 = 100$	24.07 (0.08)	7.99 (0.05)	43.57 (0.41)	33.85 (0.60)
HD	$n_1 = n_2 = 50$	25.54 (0.10)	8.77 (0.06)	35.47 (0.17)	24.27 (0.16)
	$n_1 = n_2 = 100$	24.96 (0.09)	8.38 (0.06)	34.03 (0.15)	22.86 (0.13)
SD	$n_1 = n_2 = 50$	25.71 (0.11)	8.84 (0.06)	35.70 (0.17)	24.77 (0.17)
	$n_1 = n_2 = 100$	25.07 (0.09)	8.44 (0.06)	34.01 (0.14)	23.10 (0.12)
L_1D	$n_1 = n_2 = 50$	24.78 (0.10)	8.44 (0.06)	34.77 (0.16)	23.53 (0.15)
	$n_1 = n_2 = 100$	24.40 (0.09)	8.14 (0.06)	33.17 (0.13)	22.04 (0.11)
$MD_{1/2}$	$n_1 = n_2 = 50$	24.75 (0.10)	8.43 (0.06)	31.40 (0.11)	20.18 (0.08)
	$n_1 = n_2 = 100$	24.30 (0.09)	8.12 (0.06)	30.85 (0.09)	19.93 (0.08)
$MD_{3/4}$	$n_1 = n_2 = 50$	24.38 (0.09)	8.19 (0.06)	31.52 (0.11)	20.24 (0.09)
	$n_1 = n_2 = 100$	24.17 (0.08)	8.06 (0.05)	30.90 (0.13)	19.96 (0.08)
PD	$n_1 = n_2 = 50$	25.57 (0.11)	8.76 (0.07)	33.69 (0.13)	21.80 (0.11)
	$n_1 = n_2 = 100$	24.83 (0.09)	8.33 (0.06)	32.31 (0.10)	20.82 (0.09)

Note that in these examples, the overall performance of the robust MD classifiers was better than the PD classifier (see Table 1). This is probably because of the fact that in MD and robust MD, we used the information on the homoscedastic structure of two competing populations, but for PD computation, we could not use this fact. As a result, while the MD and the robust MD classifiers were always linear, in some cases, the sample version of PD led to a nonlinear class boundary. However, this feature of PD classifier could be helpful in some cases. To demonstrate this, let us consider the example with two normal distributions having the same dispersion matrix $\mathbf{I_2}$ but different location parameters $(0, 0)$ and $(2, 2)$, but the difference is that here 10% of the class-1 observations were replaced by observations from a normal distribution with the mean $(20, 20)$ and the dispersion matrix $\mathbf{I_2}$. In the presence of these contaminating observations, LDA was affected most, and it misclassified almost half

of the observations (see the scatter plot of the data set and the class boundary estimated by LDA in Figure 1(b)). The robust MD classifiers were less affected in this case (Figure 1(c)). $MD_{1/2}$ and $MD_{3/4}$ led to average error rates of 12.71% and 12.65%, respectively, (over 500 trials) with the same standard error of 0.05%. Note that since these classifiers are linear, they were unable to correctly classify the 10% observations of class-1 located around $(20, 20)$. The PD classifier, in this case, had a significantly lower error rate of 10.78% with a standard error of 0.11%, and like the Bayes classifier (Figure 1(a)), it correctly classified these 10% observations (see Figure 1(d)). Other depth based classifiers had significantly higher misclassification rates. For L_1D, HD and SD, these error rates were 18.21%, 22.40% and 24.56%, respectively, with corresponding standard errors of 0.11%, 0.14% and 0.16%.

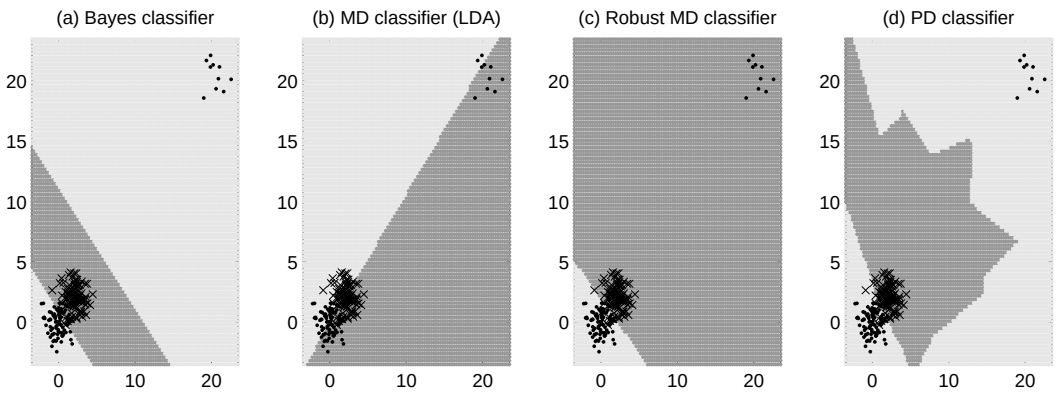

Figure 1: Class boundaries estimated by different maximum depth classifiers.

3 Generalized projection depth classifier

The PD classifier and other maximum depth classifiers described in Section 2 performs well when the priors are equal, and the population distributions differ only in their location. However, in practice, two populations may have different priors, and the population dis-

tributions may also differ in scatters and shapes. In such cases, like all maximum depth classifiers, the PD classifier may fail to perform satisfactorily, and it needs to be modified. Xia, Lin and Yang [47] proposed a modification of the PD classifier to handle such cases, but unfortunately their method is useful in the case of different scatters only. Here we propose a generalization which works under a more general set up. Here also the development of the generalized PD classifier will be motivated by elliptic symmetry of the underlying populations. Note that if the population distributions are elliptically symmetric, the optimal Bayes classifier is given by [13]

$$d_B(\mathbf{x}) = \arg\max_j \pi_j \psi_j\{D(\mathbf{x}, F_j)\},$$

where π_js are the priors of different classes, and ψ_js are appropriate transformation functions. Note that if the population distributions differ only in their location, and the elliptic distributions are unimodal, ψ_js are same for all populations, and it turns out to be a decreasing function of depth. Therefore, in such cases, when the priors are equal, the Bayes classifier turns out to be the maximum depth classifier. But, if we do not make such assumptions on the underlying distributions and their priors, in order to develop a good classifier, one needs to estimate ψ_j from the data. Unfortunately, for most of the depth functions, this function is not known explicitly, and that hinders the generalization of those depth based classifiers. Of course, it is explicitly known for MD, but its empirical version based on sample moments is not robust. So, to get a generalized robust depth based classifier, either one needs to plug-in robust estimates for the multivariate location and scatter or one needs to adopt other notions of depth, where the corresponding ψ_js are either known or can be easily estimated from the data. For PD, the function ψ_j has a nice closed form expression, and using that one arrives at the following proposition.

Proposition 1*: If the population distributions are elliptically symmetric, the optimal Bayes classifier is given by* $d_B(\mathbf{x}) = \arg\max_j \lambda_j \rho_j\{PD(\mathbf{x}, F_j)\}\{PD(\mathbf{x}, F_j)\}^{d-3}/\{1 - PD(\mathbf{x}, F_j)\}^{d-1}$, *where* $\rho_j(.)$ *is the density function of* $PD(\mathbf{x}, F_j)$*, and* λ_j *is an appropriate constant.*

Note that the above result holds for the general definition of PD, only the constant term will change depending on the choices of robust measures of univariate location and scale. In order to construct a depth based classifier, we estimate $\mathrm{PD}(\mathbf{x}, F_j)$ by its sample analog $\mathrm{PD}(\mathbf{x}, F_{jn_j})$, and the corresponding density function is estimated using the kernel density estimation technique [41]. Note that irrespective of the dimension of the measurement space, here we need one dimensional density estimation. This helps to get rid of the curse of dimensionality that one faces in high dimensional nonparametric density estimation. Density estimation using data depth and the importance of dimension reduction were also discussed in [8]. Note that the error rate of the classifier depends on the constants $\lambda_1, \lambda_2, \ldots, \lambda_J$, and we need to estimate them. In practice, in a J-class problem, one can take $\lambda_1 = 1$ and estimate $\lambda_2, \lambda_3, \ldots, \lambda_J$ by minimizing the misclassification rate of the resulting classifier.

Since the kernel method is used for the estimation of density function ρ_j, one has to find appropriate bandwidth parameters $h_j (j = 1, 2, \ldots, J)$ for the competing classes. In a two class problem, for a given value of h_1 and h_2, the kernel density estimates of the two classes are given by $\hat{\rho}_{jh_j}(\delta) = (n_j h_j)^{-1} \sum_{i=1}^{n_j} K\{h_j^{-1}(\delta - \hat{\delta}_{n_j}^{(j)}(\mathbf{x}_{ji}))\}$ for $j = 1, 2$, where $\hat{\delta}_{n_j}^{(j)}(\mathbf{x}) = PD(\mathbf{x}, F_{jn_j})$. Therefore the resulting classifier can be expressed as

$$d_{2n}(\mathbf{x}) = \begin{cases} 1 \quad \text{if} & log(r_{n_1,h_1}^{(1)}(\mathbf{x})) - log(r_{n_2,h_2}^{(2)}(\mathbf{x})) > k, \\ 2 \quad \text{otherwise} & \end{cases},$$

where $r_{n_j,h_j}^{(j)}(\mathbf{x}) = \hat{\rho}_{jh_j}(\hat{\delta}_{n_j}^{(j)}(\mathbf{x}))(\hat{\delta}_{n_j}^{(j)}(\mathbf{x}))^{d-3}/(1 - \hat{\delta}_{n_j}^{(j)}(\mathbf{x}))^{d-1}$, and $k = log(\lambda_2)$ is chosen by

minimizing the cross validation error. Clearly, the performance of the classifier $d_{2n}(\cdot)$ depends on h_1 and h_2. If h_1 and h_2 satisfy assumption (A3) (assumptions (A1)-(A3) are mentioned below), and the underlying distributions are elliptically symmetric satisfying assumptions (A1)-(A2), the error rate of the generalized PD classifier $d_{2n}(\cdot)$ converges to the Bayes risk. This is formally stated in the following theorem, and the proof is given in the Appendix. To prove this asymptotic optimality, here we assume the prior probabilities π_1 and π_2 to be known. If these priors are not known, one usually estimates them using the training sample proportions of the two classes. Since the convergence of the these estimates are relatively faster, the following theorem holds even when the estimated priors are plugged in.

Theorem 2*: Consider the following assumptions*

(A1) $f_j(\mathbf{x}) > 0$ *for all* $\mathbf{x} \in \mathbb{R}^d$ *and* $j = 1, 2$.

(A2) For $j = 1, 2$, $F_{\gamma,j}(z) = P(\gamma(\mathbf{X}) \leq z \mid \mathbf{X} \in j - th\ class)$ *is uniformly continuous in* z, *where* $\gamma(\mathbf{x}) = r^{(2)}(\mathbf{x})/r^{(1)}(\mathbf{x})$, $r^{(j)}(\mathbf{x}) = \rho_j(\delta^{(j)}(\mathbf{x}))(\delta^{(j)}(\mathbf{x}))^{d-3}/(1-\delta^{(j)}(\mathbf{x}))^{d-1}$ *and* $\delta^{(j)}(\mathbf{x}) = PD(\mathbf{x}, F_j)$.

(A3) For $j = 1, 2$, $h_j \to 0$ *and* $n_j h_j (\log n_j)^{-1} \to \infty$ *as* $n_j \to \infty$.

If (A1)-(A3) hold, and the two population distributions are elliptically symmetric, the misclassification rate of the generalized projection depth classifier $d_{2n}(\cdot)$ *converges to the Bayes risk as* $min\{n_1, n_2\} \to \infty$.

A method for density estimation and classification based on HD was proposed in [13]. But, as mentioned earlier, empirical version of HD is based on counting, and it takes some discrete values only. This discretisation leads to loss of information, and as a result, we often have poor density estimates, which have several peaks near those discrete values. Besides this,

because of the observations having zero depth, the resulting population density estimate $\hat{f}_j$ has bumps in the tail, which is not desirable. Since the empirical version of PD is continuous in $\mathbf{x}$, we do not face such problems during density estimation using PD. As a consequence, the resulting classifier usually performs much better than that derived from HD.

Theorem 2 gives us an idea about the optimal asymptotic order of the bandwidth parameters. However, in practice, when we work with a fixed sample size, one needs to estimate these bandwidths from the data. Instead of using bandwidths that minimize the estimated mean integrated squared errors [41] of kernel density estimates, in classification problems, it is better to choose them by minimizing the estimated error rate of the resulting classifier [11]. Here, we used leave-one-out cross-validation method to estimate this error rate and chose the bandwidths h_1, h_2 and the constant λ_2 by minimizing this estimated error. Since the optimal smoothing parameter is supposed to be proportional to the population dispersion, to reduce the computational cost, we chose $h_1 = (s_1/s_2)h_2$, where s_j $(j = 1, 2)$ is a dispersion measure of the estimated depth functions $\hat{\delta}^{(j)}_{n_j}(\mathbf{x}_{j1}), \hat{\delta}^{(j)}_{n_j}(\mathbf{x}_{j2}), \ldots, \hat{\delta}^{(j)}_{n_j}(\mathbf{x}_{jn_j})$. To make it robust, here we used sample quartile deviation to compute s_1 and s_2. Clearly, the estimated misclassification rate turns out to be a function of h_2 and λ_2. In order to estimate λ_2 or $k = log(\lambda_2)$, we compute $r^{(i)}_{n_i,h_i}(\mathbf{x}_{jl}) = \hat{\rho}^{*}_{ih_i}(\hat{\delta}^{(i)}_{n_i}(\mathbf{x}_{jl}))(\hat{\delta}^{(i)}_{n_i}(\mathbf{x}_{jl}))^{d-3}/(1 - \hat{\delta}^{(i)}_{n_i}(\mathbf{x}_{jl}))^{d-1}$ for $i, j = 1, 2$ and $l = 1, 2, \ldots, n_j$, where $\hat{\rho}^{*}$ stands for the usual kernel density estimate for $j \neq i$, and it stands for the leave one out density estimate for $j = i$. The constant k can be searched over the order statistics of $log(r^{(1)}_{n_1,h_1}(\mathbf{x}_{jl})) - log(r^{(2)}_{n_2,h_2}(\mathbf{x}_{jl}))$ $(j = 1, 2, l = 1, 2, \ldots, n_j)$ to minimize the cross validation error rate. Clearly this choice of k depends on h_2. We used different choices of h_2 over a suitable range to compute the cross-validation error rate and chose the one that

leads to the lowest error rate. Due to the stepwise nature of the cross validation error rate, sometimes multiple values of h_2 lead to the lowest error rate. Following [11], we have chosen the maximum of the optimizers in such cases.

Note that similar generalization is also possible for depth based classification using MD and robust MD. Under appropriate conditions, the robust estimates of scatter parameters computed using MCD or other methods converge to a scalar multiple of the population scatter matrix. Since this scalar quantity depends on the underlying population, it needs to be estimated it from the data. However, from the above discussion, it is clear that whatever be that scalar, the form of the Bayes classifier remains the same as in Proposition 1. So, the classification method discussed above can also be adopted to develop generalized robust MD classifier, and its asymptotic optimality can be shown following the proof of Theorem 2.

For classification problems with more than two classes, the bandwidths $h_1, h_2, \ldots, h_J$ and the constants $\lambda_1, \lambda_2, \ldots, \lambda_J$ can be chosen in a similar way. But it becomes computationally difficult to minimize the cross-validation error rate w.r.t. several parameters. To reduce this computational cost, we adopt a different strategy. Here, we look at this J-class problem as $\binom{J}{2}$ distinct binary classification problems. Taking each pair of classes at a time, we perform binary classification, and the results of all pairwise classifications are aggregated to arrive at the final decision. One can use majority voting [10] or pairwise coupling [16] method for this aggregation. In this article, for computational simplicity, we have used the voting method, where ties are solved arbitrarily. Following the arguments of the proof of Theorem 2, under similar regularity conditions, one can show the convergence of the error rate of the generalized PD classifier to the Bayes risk even for this multi-class problem.

3.1 Performance on simulated data sets

Here we analyze some simulated examples to show the utility of the proposed generalization. As we have discussed before, the maximum depth classifiers discussed in Section 2 work well when the population distributions have same prior, they are nearly elliptic, and they have the same distributional structure except for location. However, in practice, the prior probabilities of different classes may be different, and the competing population distributions may also have different scatters and shapes. In such cases, the maximum depth classifiers may lead to drastically poor performance, but the generalized PD classifier can perform much better. Our analysis of simulated datasets makes it more transparent.

We begin with the same examples with normal and Cauchy distributions considered in Section 2, where the two competing populations differ only in their location. But here instead of taking $\pi_1 = \pi_2$, we consider three different choices for $\pi_1 = 0.5$, 0.6 and 0.7. Note that for $\pi_1 = 0.5$, results of different maximum depth classifiers have already been reported in Table 1. As compared to these results, generalized depth based classifiers had marginally higher error rates in some cases, while in some other cases, the error rates remained almost the same. It is quite expected as we are not using the information about the location shift model, and in addition to depth, here one needs to estimate the density functions as well. But, unlike maximum depth classification, here one can modify the classification rule as the prior changes. So, irrespective of the prior probabilities, when the maximum depth classifiers lead to the same error rate, generalized depth based classifier can reduce the error rate significantly when the priors are different. As a result, for $\pi_1 = 0.7$ and even for $\pi_1 = 0.6$, the error rates yielded by the generalized PD classifier were found to be much lower than

that obtained using the PD classifier in Section 2. We observed the same phenomenon for the generalized MD and the generalized HD classifiers as well. One should also notice that the generalized PD and the generalized robust MD classifiers worked much better than the corresponding HD version proposed in [13]. In majority of the cases, they significantly outperformed the generalized HD classifier.

Table 2 : Error rates of different generalized depth based classifiers and their standard errors.

			$\mu = 1$			$\mu = 2$		
			$\pi_1 = 0.5$	$\pi_1 = 0.6$	$\pi_1 = 0.7$	$\pi_1 = 0.5$	$\pi_1 = 0.6$	$\pi_1 = 0.7$
Normal distribution	Bayes risk $\longrightarrow$		23.98	23.10	20.41	7.86	7.66	7.00
	LDA	$n_1 = n_2 = 50$	24.18 (.09)	23.27 (.09)	20.73 (.08)	8.05 (.06)	7.86 (.05)	7.21 (.05)
		$n_1 = n_2 = 100$	24.07 (.08)	23.21 (.08)	20.55 (.08)	7.99 (.05)	7.73 (.05)	7.07 (.05)
	HD	$n_1 = n_2 = 50$	26.66 (.13)	26.27 (.15)	24.31 (.13)	9.09 (.07)	8.94 (.07)	8.68 (.09)
		$n_1 = n_2 = 100$	25.69 (.10)	24.98 (.10)	22.61 (.10)	8.52 (.06)	8.37 (.06)	7.90 (.06)
	MD	$n_1 = n_2 = 50$	25.71 (.11)	24.89 (.11)	22.63 (.11)	8.81 (.07)	8.62 (.07)	7.96 (.07)
		$n_1 = n_2 = 100$	24.99 (.09)	24.23 (.09)	21.80 (.09)	8.40 (.06)	8.17 (.06)	7.57 (.06)
	$MD_{1/2}$	$n_1 = n_2 = 50$	26.50 (.13)	25.89 (.13)	23.47 (.13)	9.23 (.08)	9.05 (.08)	8.29 (.07)
		$n_1 = n_2 = 100$	25.32 (.10)	24.52 (.09)	22.17 (.09)	8.58 (.06)	8.31 (.06)	7.66 (.06)
	$MD_{3/4}$	$n_1 = n_2 = 50$	25.92 (.12)	25.26 (.12)	22.89 (.12)	9.03 (.07)	8.82 (.07)	8.15 (.07)
		$n_1 = n_2 = 100$	25.09 (.10)	24.39 (.09)	22.00 (.09)	8.51 (.06)	8.26 (.06)	7.61 (.06)
	PD	$n_1 = n_2 = 50$	26.33 (.13)	25.72 (.14)	23.13 (.13)	9.14 (.08)	8.97 (.07)	8.26 (.07)
		$n_1 = n_2 = 100$	25.40 (.10)	24.59 (.10)	21.87 (.09)	8.59 (.06)	8.39 (.06)	7.76 (.06)
Cauchy distribution	Bayes risk $\longrightarrow$		30.40	28.88	25.01	19.60	18.77	16.68
	LDA	$n_1 = n_2 = 50$	42.61 (.43)	40.28 (.05)	30.65 (.04)	32.27 (.60)	38.40 (.17)	30.49 (.05)
		$n_1 = n_2 = 100$	43.57 (.41)	40.25 (.03)	30.32 (.02)	33.85 (.60)	39.89 (.07)	30.39 (.03)
	HD	$n_1 = n_2 = 50$	36.54 (.18)	35.35 (.19)	31.23 (.19)	25.25 (.17)	24.54 (.18)	22.45 (.19)
		$n_1 = n_2 = 100$	34.49 (.14)	33.28 (.14)	29.32 (.13)	23.12 (.12)	22.54 (.12)	20.55 (.11)
	MD	$n_1 = n_2 = 50$	38.91 (.26)	37.18 (.22)	32.03 (.17)	27.06 (.28)	26.49 (.28)	24.67 (.26)
		$n_1 = n_2 = 100$	38.88 (.25)	36.72 (.19)	30.70 (.10)	26.75 (.27)	26.09 (.26)	23.92 (.21)
	$MD_{1/2}$	$n_1 = n_2 = 50$	33.51 (.13)	32.80 (.15)	29.77 (.18)	21.86 (.11)	21.37 (.11)	19.78 (.12)
		$n_1 = n_2 = 100$	32.28 (.10)	31.17 (.11)	27.71 (.13)	20.95 (.09)	20.50 (.09)	18.70 (.09)
	$MD_{3/4}$	$n_1 = n_2 = 50$	33.68 (.13)	32.82 (.15)	29.91 (.18)	21.96 (.11)	21.54 (.11)	19.75 (.12)
		$n_1 = n_2 = 100$	32.43 (.11)	31.21 (.11)	27.77 (.13)	21.00 (.09)	20.54 (.09)	18.58 (.09)
	PD	$n_1 = n_2 = 50$	33.37 (.13)	32.51 (.14)	28.95 (.16)	21.70 (.10)	21.35 (.11)	19.56 (.11)
		$n_1 = n_2 = 100$	32.26 (.10)	31.25 (.10)	27.66 (.11)	20.94 (.09)	20.45 (.09)	18.86 (.10)

Next, we consider some examples where the two competing populations have same prior probability, but unlike the previous cases, they have no difference in their location. Here also, we consider some examples with normal and Cauchy distributions. In each case, the competing populations are chosen to have the same location parameters $(0, 0)$ but different

Table 3 : Error rates of different generalized depth based classifiers and their standard errors.

		Normal		Cauchy	
		σ^2=4	σ^2=9	σ^2=4	σ^2=9
Bayes risk $\longrightarrow$		26.37	16.22	37.00	30.15
QDA	$n_1 = n_2$ =50	27.34 (0.09)	17.00 (0.08)	47.43 (0.17)	44.19 (0.24)
	$n_1 = n_2$ =100	26.84 (0.08)	16.51 (0.07)	48.25 (0.13)	46.97 (0.27)
HD	$n_1 = n_2$ =50	36.57 (0.26)	28.92 (0.31)	43.34 (0.17)	37.06 (0.20)
	$n_1 = n_2$ =100	31.88 (0.19)	24.94 (0.25)	41.28 (0.14)	34.37 (0.15)
MD	$n_1 = n_2$ =50	30.93 (0.14)	19.61 (0.12)	42.66 (0.17)	36.23 (0.17)
	$n_1 = n_2$ =100	29.22 (0.11)	18.21 (0.09)	42.32 (0.14)	35.57 (0.17)
$MD_{1/2}$	$n_1 = n_2$ =50	33.08 (0.17)	21.34 (0.15)	42.55 (0.17)	35.17 (0.17)
	$n_1 = n_2$ =100	30.16 (0.12)	19.04 (0.09)	41.32 (0.14)	33.78 (0.11)
$MD_{3/4}$	$n_1 = n_2$ =50	31.84 (0.15)	20.46 (0.13)	42.03 (0.16)	34.50 (0.14)
	$n_1 = n_2$ =100	29.43 (0.12)	18.45 (0.09)	41.37 (0.13)	33.77 (0.12)
PD	$n_1 = n_2$ =50	30.89 (0.14)	19.63 (0.13)	42.18 (0.18)	34.36 (0.15)
	$n_1 = n_2$ =100	28.86 (0.11)	17.89 (0.09)	40.44 (0.13)	32.63 (0.12)

scatter matrices $\mathbf{I_2}$ and $\sigma^2\mathbf{I_2}$, respectively for the first and the second populations. We used two different values of $\sigma^2 = 4$ and 9, and the error rates of different generalized depth based classifiers are reported in Table 3. Clearly, the optimum class boundary is quadratic in these problems. Therefore, to facilitate the comparison, we have reported the error rates of usual quadratic discriminant analysis (QDA). For further comparison, Bayes error rates are reported as well. Since there was no difference between the location of the two populations, all maximum depth classifiers led to an error rate of almost 50%, but the generalized versions worked well. In the case of normally distributed data, as expected, QDA led to the best performance, but the performance of the generalized PD classifier was fairly satisfactory. Its overall performance was the best among the generalized depth based classifiers, and it had an improvement of 10-30% over the error rates of the corresponding HD classifier. Like LDA, in the case of Cauchy distribution, QDA misclassified almost half of the test set observations. But, in this case, the generalized PD classifier worked well, and its overall performance was better than the generalized MD and HD classifiers, especially when $n_1 = n_2 = 100$.

Next, we consider some cases, where the location and the scatter parameters of the two populations are same, but they differ in shapes. We start with a classification problem between a standard multivariate normal and a standard multivariate Cauchy distributions. Here also the optimum class boundary is quadratic. So, along with the Bayes error, we have reported the error rates of QDA. In this example, the maximum depth classifiers again yielded error rates close to 50%, but the generalized versions worked well. QDA led to the best performance in this example, but the error rates of the generalized depth based classifiers were very close. Once again, the generalized MD and PD classifiers performed better than HD. One should also notice that with the increasing sample size, while the error rate of QDA remained almost the same, that of the depth based methods reduced significantly. So, for sufficiently large sample, these methods are expected to outperform QDA.

Next, we consider a classification problem, where each of the two classes are standard bivariate normal but one of them is truncated to have $\mathbf{x}$ with $\|\mathbf{x}\| \geq 4$. In this case, QDA and the generalized depth based classifiers, except $\text{MD}_{1/2}$, performed well, with MD and HD versions having an edge. For $n_1 = n_2 = 50$, they had much better performance than the PD version, but this difference was smaller for $n_1 = n_2 = 100$. We carried out the same experiment also with Cauchy distributions. In this case, the maximum depth classifiers and QDA had error rates close to 50%, but the generalized depth based classifiers had much lower misclassification rates. In this example, the generalized PD classifier performed significantly better than all other depth based classification methods considered in this article.

Finally, we consider an example where one population is bivariate normal with mean $(0, 0)$ and dispersion matrix $25\mathbf{I_2}$, and the other one is an equal mixture of two bivariate normal

Table 4 : Error rates of different generalized depth based classifiers and their standard errors.

	Bayes risk ↓		QDA	HD	MD	$MD_{1/2}$	$MD_{3/4}$	PD
Normal vs.	33.57	$n_1 = n_2$ =50	34.93 (0.09)	40.46 (0.21)	36.93 (0.14)	38.62 (0.16)	37.84 (0.15)	38.40 (0.15)
Cauchy		$n_1 = n_2$ =100	35.12 (0.09)	38.88 (0.18)	35.80 (0.11)	36.63 (0.13)	36.17 (0.12)	36.54 (0.12)
Normal vs.	6.77	$n_1 = n_2$ =50	15.52 (0.17)	13.42 (0.13)	13.81 (0.20)	23.68 (0.22)	15.07 (0.19)	17.02 (0.19)
Trun. normal		$n_1 = n_2$ =100	13.16 (0.13)	11.27 (0.07)	10.72 (0.09)	20.73 (0.18)	11.40 (0.11)	13.29 (0.14)
Cauchy vs.	22.36	$n_1 = n_2$ =50	46.07 (0.21)	33.05 (0.21)	35.06 (0.22)	30.80 (0.16)	30.63 (0.15)	28.91 (0.15)
Trun. Cauchy		$n_1 = n_2$ =100	47.50 (0.15)	30.09 (0.18)	34.51 (0.20)	28.36 (0.12)	28.34 (0.12)	26.43 (0.11)
Normal vs.	21.75	$n_1 = n_2$ =50	46.84(0.17)	38.87 (0.22)	36.56 (0.24)	34.99 (0.20)	38.40 (0.20)	27.82 (0.17)
Mix. normal		$n_1 = n_2$ =100	45.81(0.10)	35.28 (0.18)	32.51 (0.23)	30.65 (0.16)	34.31 (0.19)	24.74 (0.11)
Cauchy vs.	38.28	$n_1 = n_2$ =50	50.00 (0.20)	46.51 (0.18)	45.78 (0.16)	45.96 (0.14)	46.73 (0.13)	43.74 (0.15)
Mix. Cauchy		$n_1 = n_2$ =100	50.00 (0.19)	43.95 (0.15)	45.19 (0.14)	45.48 (0.15)	45.69 (0.15)	41.99 (0.13)

distributions having the same mean $(0, 0)$ but different dispersion matrices $\mathbf{I_2}$ and $100\mathbf{I_2}$. In this case, the generalized PD classifier performed much better than all other classifiers, and it had error rates close to the Bayes risk. For $n_1 = n_2 = 100$, when it yielded an error rate of 24.74 %, error rates of all other classifiers were higher than 30%. We observed similar results for $n_1 = n_2 = 50$ as well. This superiority of the generalized PD classifier was also evident when we carried out the same experiment with Cauchy distributions.

4 Multi-scale analysis

To construct the generalized depth based classifiers, one needs to find the smoothing parameters (bandwidths) involved in kernel density estimation. Since the performance of the resulting classifier depends on these smoothing parameters, they have to be chosen judiciously. In Section 3, we have used the cross-validation method for this purpose. But this way of choosing only one bandwidth pair (h_1, h_2) brings in the model uncertainty. Moreover, one should notice that in addition to depending on the training sample, a good choice of the smoothing parameter depends on the specific observation to be classified. A fixed level of

smoothing may not work well in all parts of the measurement space. Therefore, instead of working with a fixed (h_1, h_2), it would be more useful to simultaneously study the classification results for different levels of smoothing in an appropriate range. Results obtained at multiple scales of smoothing can be aggregated to arrive at a new classifier, which we call the multi-scale classifier. The usefulness of this multi-scale classification has been discussed in the literature both for kernel discriminant analysis and nearest neighbor classification [21, 22, 14, 15]. One popular way to aggregate the results of these different classifiers is to take the weighted average of the estimated posterior probabilities. Popular ensemble methods like bagging [1] and boosting [38] also adopt similar methods. Note that for fixed (h_1, h_2), we classify an observation $\mathbf{x}$ to Class-1 if $\xi_{n,h_1,h_2}(\mathbf{x}) = \log(r^{(1)}_{n_1,h_1}(\mathbf{x})) - \log(r^{(2)}_{n_2,h_2}(\mathbf{x})) - k > 0$, where k is chosen by minimizing the cross validation error. So, $e^{\xi_{n,h_1,h_2}(\mathbf{x})}$ gives an estimate of the ratio of two population densities. From this, one can obtain the estimated posterior for Class-1, which is given by $\hat{p}_{n,h_1,h_2}(1|\mathbf{x}) = e^{\xi_{n,h_1,h_2}(\mathbf{x})}/(1 + e^{\xi_{n,h_1,h_2}(\mathbf{x})})$. We aggregate these posterior probabilities obtained at different values of (h_1, h_2) to arrive at the final classifier

$$d_{3n}(\mathbf{x}) = \text{argmax}_{j=1,2}\ p^*_n(j|\mathbf{x})\ \text{, where}\ \ p^*_n(j|\mathbf{x}) = \textstyle\sum_{h_1,h_2 \in H} w_{h_1,h_2}\ \hat{p}_{n,h_1,h_2}(j|\mathbf{x}),$$

for w_{h_1,h_2} being the weight function assigned to the classifier that uses h_1 and h_2 as the bandwidths of the two classes. Clearly, this aggregated classifier depends on the bandwidth range $H = [h^l_1, h^u_1] \times [h^l_2, h^u_2]$ and the weight function w. Interestingly, if the upper and the lower bounds (h^u_j and h^l_j) of h_j (for $j = 1, 2$) satisfy assumption (A3), irrespective of the choice of weight function, the error rate of $d_{3n}(\cdot)$ converges to the Bayes risk.

Theorem 3 *: Assume that for $j = 1, 2$, h^u_j and h^l_j satisfy (A3). Also assume that the two population distributions are elliptically symmetric, and they satisfy (A1)-(A2). Then the*

misclassification rate of the multi-scale generalized projection depth classifier $d_{3n}(\cdot)$ converges to the Bayes risk as $min\{n_1, n_2\} \to \infty$.

From the above result, it is quite clear that the large sample performance of the multi-scale classifier is not very sensitive to the choice of the weight function w. However, in practice, when we deal with a sample of fixed size, one has to choose H and w appropriately. Instead of using equal weights for all classifiers, naturally one would like to use higher weights w_{h_1,h_2} for classifiers having lower misclassification rates (Δ_{h_1,h_2}), and the weight should decrease gradually as the error rate increases. Boosting [38] also adopts the same idea and uses different weights for different classifiers based on their corresponding error rates. However, in [15] the authors argued that the log function used in boosting fails to appropriately weigh down the poor classifiers resulting from poor choices of h_1 and h_2. Instead, they proposed a Gaussian weight function that decreases at an exponential rate. In this article, for our data analysis, following [15], we estimated Δ_{h_1,h_2} by cross validation method and used the weight function $w_{h_1,h_2} = \exp\left[-\frac{1}{2}\frac{(\hat{\Delta}_{h_1,h_2}-\hat{\Delta}_0)^2}{\hat{\Delta}_0(1-\hat{\Delta}_0)/N}\right] I[\hat{\Delta}_{h_1,h_2} \leq \min\{\pi_1, \pi_2\}]$, where $N = n_1+n_2$ and $\hat{\Delta}_0 = \min_{h_1,h_2} \hat{\Delta}_{h_1,h_2}$. Note that $\hat{\Delta}_0$ and $\hat{\Delta}_0(1-\hat{\Delta}_0)/N$ can be viewed as the estimates for the mean and the variance of the empirical misclassification rate of the best single-scale generalized PD classifier, when such a classifier is used to classify N independent observations. Also notice that $\min\{\pi_1, \pi_2\}$ is the error rate of the trivial classifier that classifies all observations to the class having larger prior. If the classifier with bandwidth pair (h_1, h_2) is worse than that, the weighing scheme ignores it by putting zero weight. For the choice H, we have followed the method based on quantiles of the pairwise distances as described in [15]. However, instead of considering all values of (h_1, h_2) in H, for computational simplicity, we have considered

100 equidistant values of (h_1, h_2) satisfying $h_1 = (s_1/s_2)h_2$, where s_1 and s_2 have the same meaning as in Section 3. Though this choice of the weight function and the bandwidth range is somewhat subjective, it yielded good results in our experiments.

4.1 Performance on simulated data sets

The primary focus of this sub-section is to show the utility of multi-scale analysis in generalized depth based classification. For this, we have analyzed the same simulated data sets used for single-scale classification. However, instead of presenting all results in another table, for better visualization, following the idea of [9], here we used the notion of efficiency to compare the performance of different single-scale and multi-scale classifiers. Since the generalized HD classifier had much higher error rates compared to generalized PD and MD classifiers, we did not consider it for this comparison. Note that we have analyzed 13 simulated data sets in this paper. For each data set, we define the efficiency of the t-th classifier as $\eta_t = \{\min_t e_t\}/e_t$, where e_t is the error rate of the t-th classifier, and the minimization is done over all classifiers considered for comparison (i.e. single-scale and multi-scale versions of MD, $\text{MD}_{1/2}$, $\text{MD}_{3/4}$, PD). Note that on each data set, the best classifier has $\eta_t = 1$ and it is less than 1 for all other classifiers. Small value of η_t shows the lack of robustness of the t-th classifier. We computed these efficiencies for different single-scale and multi-scale classifiers on all these 13 simulated data sets when 100 observations from each class constituted the training sample, and the results are summarized using box-plots in Figure 2. This figure clearly shows the importance of multi-scale classification. For all generalized depth based classifiers, overall performance of the multi-scale methods was better than their single-scale

counter parts. Among different depth based classifiers, MD shows higher dispersion in efficiency because of its lack of robustness. The multi-scale version of PD had the best overall performance, followed by its single-scale version. The performance of $MD_{1/2}$ and $MD_{3/4}$ was also comparable. We observed similar results when we carried out the same experiment taking 50 observations from each class in the training sample.

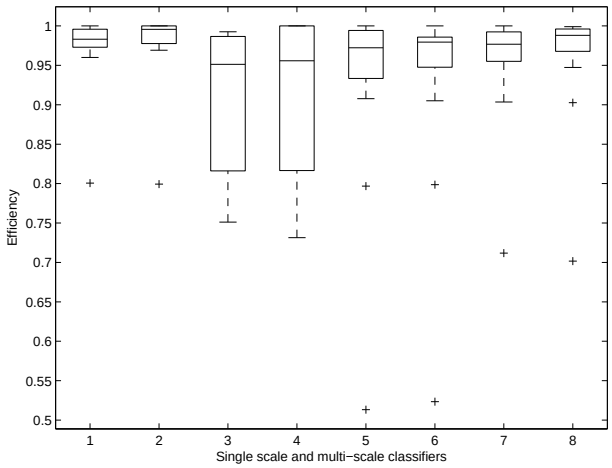

Figure 2: Efficiencies of different single-scale and multi-scale classifiers : 1 PD-SS, 2 PD-MS, 3 MD-SS, 4 MD-MS, 5 $MD_{1/2}$-SS, 6 $MD_{1/2}$-MS, 7$MD_{3/4}$-SS, 8 $MD_{3/4}$-MS.

5 Analysis of benchmark data sets

Now, we investigate the performance of different depth based classifiers on some well known benchmark data sets. Except for the vowel data, all these data sets are taken from CMU data archive (http://www.statlib.cmu.edu). For the synthetic data and the vowel data, the training and the test sets are well specified. In these cases, we report the test set error rates of different classifiers. In other data sets (i.e. the glass data and the biomedical data),

we formed these training and test sets by randomly partitioning the data. This random partitioning was done 500 times to generate different training and test sets, and the average test set misclassification rates (over these 500 trials) of different depth based classifiers are reported here (see Table 5) along with their corresponding standard errors. In each case, the training and the test sets are formed such that the proportion of different classes in these two sets are as close as possible. Along with the error rates of single-scale (SS) and multi-scale (MS) generalized depth based classifiers, we also report the performance of some standard parametric (LDA and QDA) and nonparametric (kernel discriminant analysis (KDA) and nearest neighbor classification (NN)) classifiers to facilitate the comparison. Throughout this section, training sample proportions of different classes are taken as their prior probabilities. In the case of glass data and biomedical data, since the dimension of the measurement vector was greater than 2, it was not computationally feasible to adopt the grid search method used in simulation studies. Instead, one can use the Nedler-Mead algorithm (available in R, see e.g., [46] or the gradient descent algorithm [39] for computing PD. The later one is computationally much faster, but it has the problem of getting stuck at poor local minima. As a remedy, we ran the algorithm several times starting from different initial points and chose the one that gave the best result in the training set. We tried both these algorithms, but the error rates were almost the same. Therefore, to avoid repetition, here we have reported the result for Nedler-Mead algorithm only.

Synthetic data was generated and extensively analyzed by Ripley [34]. Here each of the two classes is an equal mixture of two bivariate normal distributions differing only in their location. There are 250 observations in the training set and 1000 observations in the test

set, which are equally distributed among the two classes. Since the prior probabilities of these two classes are equal, we tried out both maximum depth and generalized depth based classification. For maximum depth classification, error rates of PD (10.3 %), MD (10.8 %), $MD_{1/2}$ (11.5%) and $MD_{3/4}$ (11.50 %) were quite close, but HD (12.8%) and SD (13.8%) had relatively higher error rates. In this data set, the underlying distributions are neither elliptically symmetric nor they satisfy any location shift model. But instead of that, error rates of the maximum depth classifiers were comparable to that of the nonparametric methods like KDA (11.0%) and k-NN (11.7%). For all depth functions except MD, the performance of generalized depth based classifiers was better than the maximum depth classifier. In this data set, the single-scale version of the generalized PD classifier had the best error rate.

Vowel data was created by Petersen and Barney by a spectrographic analysis on vowels, and a detailed description of this data set is given in [33]. There were 67 speakers who uttered 10 different words starting with h followed by a vowel and then followed by d, and the two lowest resonant frequencies of their vocal tracts were noted for ten different vowels. In the training and the test sets, there are 338 and 333 observations, respectively. In this data set, k-NN led to the best error rate (17.75%), but the performance of all other classifiers, except LDA (error rate 25.26%) and generalized HD (error rate 35.73%) was also competitive.

Glass data has been previously analyzed in many articles [34, 15]. Though there are originally 214 observations in this data set, 146 (70 + 76) of them were from two bigger classes. Here we have considered those two classes only. We have taken 50 observations from each class to constitute the training sample, while the remaining observations were used as test cases. There are nine variables in the data set, but four of them had almost all values equal to zero.

Considering them to be of less important for classification, we carried out our analysis with the remaining five. Traditional parametric classifiers, LDA and QDA, performed very poorly here. So, we may suspect the lack of normality in the data. KDA and k-NN, which do not make any parametric assumption, had better error rates in this data set. Generalized PD and robust MD classifiers performed much better than parametric methods, and their error rates were close to that of nonparametric classifiers. Note that these depth based methods assume ellipticity but not normality of the underlying distributions. Results on this data set shows the advantage of working with a broader class.

Biomedical data contains four different measurements on blood samples taken from normal people and also from carriers of a genetic disorder. In this data set, there are 15 observations with missing values. We ignored them and carried out our analysis using the remaining 194 cases (127 normal, 67 carriers). Here, among the parametric and non-parametric classifiers, QDA yielded the best performance. While KDA and k-NN had error rates around 16-17%, QDA had an average error rate of 12.57%. This gives the indication that the Gaussian model may fit the data well. So, as expected, because of the validity of underlying model assumptions, generalized depth based classifiers worked well. In fact, all generalized depth based classifiers had significantly better performance than nonparametric methods. In this example, generalized MD and PD classifiers had error rates even smaller than that of QDA.

One should also notice that in these benchmark data sets, the overall performance of multi-scale classifiers were better than single-scale methods. It becomes more evident in the case of glass data and biomedical data, when the error rates are computed over 500 simulations. In those two cases, all multi-scale methods outperformed their single-scale counter parts.

Table 5 : Misclassification rates (in %) of different parametric, nonparametric and generalized depth based classifiers and their standard errors in real data sets.

Method	Synthetic data	Vowel data	Glass data	Biomedical data
LDA	10.80	25.26	30.59 (0.25)	15.66(0.14)
QDA	10.20	19.83	36.13 (0.26)	12.57 (0.12)
k-NN	11.70	17.75	22.88 (0.24)	17.88 (0.15)
KDA	11.00	19.85	22.07 (0.23)	16.82 (0.14)
HD	12.00	35.73	33.93 (0.29)	14.11 (0.14)
MD(SS)	13.00	20.75	26.59 (0.25)	12.44 (0.13)
MD(MS)	11.60	20.70	26.14 (0.25)	12.04 (0.12)
$MD_{1/2}$(SS)	11.00	19.22	26.02 (0.29)	14.64 (0.14)
$MD_{1/2}$(MS)	10.10	19.23	26.08 (0.28)	14.58 (0.14)
$MD_{3/4}$(SS)	10.30	19.22	24.92 (0.25)	14.25 (0.13)
$MD_{3/4}$(MS)	10.40	19.23	24.43 (0.25)	14.03 (0.14)
PD(SS)	10.00	20.80	25.70 (0.34)	12.37 (0.14)
PD(MS)	10.50	21.56	25.24 (0.33)	12.18 (0.13)

6 Concluding remarks

This paper investigates possible applications of PD in supervised classification. Like robust MD, the use of PD makes the classifier robust against contaminating observations and outliers generated from heavy-tailed distributions. Unlike the usual version of L_1D, PD does not suffer from lack of affine invariance. Moreover, because of the continuity of its empirical version, it usually performs better than HD and SD, which are based on counting. Another major advantage of using PD is its simple relationship with Mahalanobis distance, and because of this relationship, maximum depth classifiers can be generalized easily. The resulting generalized PD classifier performs well for a general class of classification problems. While usual parametric methods like LDA and QDA work well under the normality of underlying distributions, the depth based methods discussed here cater a more general class of parametric models. Again, unlike usual nonparametric classifiers, here we do not need to go for multivariate density estimation, and this helps to get rid of the curse of dimensionality. So if we have a small training set in higher dimensions, depth based methods are expected to

outperform them when the data clouds are nearly elliptic, which is not rare in practice.

The multi-scale method proposed here is simple and easy to implement. It provides the flexibility of considering the results for different scales of smoothing simultaneously. While small scales of smoothing take care of the local nature of the density function and the class boundary, large scales capture the global pattern. Aggregating these two important features, one can expect to have better performance of the resulting classifier. Using several simulated and benchmark data sets, we have amply demonstrated it in this article.

Appendix : Proofs

Proof of Theorem 1 : The average misclassification rate of the maximum depth classifier based on PD is given by $\Delta_n = \sum_{j=1}^{J} \pi_j P\{d_{1n}(\mathbf{X}) \neq j | \mathbf{X} \in j\text{-}th\ class\}$. Note that, if the population distributions are elliptically symmetric, and they differ only in their location parameter, the population PD based classifier is the Bayes classifier when $\pi_j = 1/J$ for all $j = 1, 2, \ldots, J$. Thus, in this case, we have

$$|\Delta_n - \Delta| \leq \frac{1}{J} \sum_{j=1}^{J} \int \Big| \prod_{i=1, i\neq j}^{J} I\left\{ \frac{\mathrm{PD}(\mathbf{x}, F_{jn_j})}{\mathrm{PD}(\mathbf{x}, F_{in_i})} > 1 \right\} - \prod_{i=1, i\neq j}^{J} I\left\{ \frac{\mathrm{PD}(\mathbf{x}, F_j)}{\mathrm{PD}(\mathbf{x}, F_i)} > 1 \right\} \Big| \, f_j(\mathbf{x}) d\mathbf{x},$$

where $\mathrm{PD}(\mathbf{x}, F_{jn_j})$ and $\mathrm{PD}(\mathbf{x}, F_j)$ are the empirical and the population PD of $\mathbf{x}$ w.r.t. the j-th population, respectively, and Δ denotes the Bayes risk. Now, under elliptic symmetry of $f_j(\mathbf{x})$, PD satisfies all the conditions required for almost sure uniform convergence of the empirical depth function to its population analog (see e.g., Zuo & Serfling, 2000b). So, we have, $\sup_{x \in \mathbb{R}^d} |\mathrm{PD}(x, F_n) - \mathrm{PD}(x, F)| \overset{a.s.}{\to} 0$. Now, the convergence of Δ_n to Δ follows from the Dominated Convergence Theorem (DCT). □

Lemma 1 : If the population distribution F is elliptically symmetric with the location parameter μ and the scale parameter Σ, we have $\{(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}^{1/2} = C_F.O(\mathbf{x},F)$, where C_F is a constant.

Proof of Lemma 1 : Since $\boldsymbol{\alpha}'\mathbf{X}$ is symmetric about $\boldsymbol{\alpha}'\boldsymbol{\mu}$, $\mu_F(\boldsymbol{\alpha}'\mathbf{X}) = \boldsymbol{\alpha}'\boldsymbol{\mu}$ for any $\boldsymbol{\alpha} \in \mathbb{R}^d$. So, $O(\mathbf{x},F) = \max_{\boldsymbol{\alpha}} \left\{ \frac{|\boldsymbol{\alpha}'\mathbf{x} - \mu_F(\boldsymbol{\alpha}'\mathbf{X})|}{\sigma_F(\boldsymbol{\alpha}'\mathbf{X})} \right\} = \max_{\boldsymbol{\alpha}} \left\{ \frac{|\boldsymbol{\alpha}'\mathbf{x} - \boldsymbol{\alpha}'\boldsymbol{\mu}|}{sd_F(\boldsymbol{\alpha}'\mathbf{X})} . \frac{sd_F(\boldsymbol{\alpha}'\mathbf{X})}{\sigma_F(\boldsymbol{\alpha}'\mathbf{X})} \right\}$. Now, define $\mathbf{Y} = \Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu})$, which is spherically distributed. So for any $\boldsymbol{\alpha} \in \mathbb{R}^d$, $\boldsymbol{\alpha}'\mathbf{Y} \stackrel{d}{=} ||\boldsymbol{\alpha}||Y_1$ [7], and $\boldsymbol{\alpha}'\mathbf{X} = \boldsymbol{\alpha}'\boldsymbol{\mu} + \boldsymbol{\alpha}'\Sigma^{1/2}\mathbf{X} = \mu_{\boldsymbol{\alpha}} + l'_{\boldsymbol{\alpha}}\mathbf{Y} \stackrel{d}{=} \mu_{\boldsymbol{\alpha}} + ||l_{\boldsymbol{\alpha}}||Y_1$, where $\mu_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}'\boldsymbol{\mu}$ and $l_{\boldsymbol{\alpha}} = \Sigma^{1/2}\boldsymbol{\alpha}$. Now, $\sigma_F(\boldsymbol{\alpha}'\mathbf{X}) = ||l_{\boldsymbol{\alpha}}||\sigma_F(Y_1)$ and $sd_F(\boldsymbol{\alpha}'\mathbf{X}) = ||l_{\boldsymbol{\alpha}}||sd_F(Y_1) \Rightarrow \frac{sd(\boldsymbol{\alpha}'\mathbf{X})}{\sigma(\boldsymbol{\alpha}'\mathbf{X})} = \frac{sd(Y_1)}{\sigma(Y_1)} = 1/C_F$(say). Since C_F is free of $\boldsymbol{\alpha}$, using the fact $\max_{\boldsymbol{\alpha}}\{|\boldsymbol{\alpha}'\mathbf{x} - \boldsymbol{\alpha}'\boldsymbol{\mu}|/sd(\boldsymbol{\alpha}'\mathbf{X})\} = \{(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}^{1/2}$, the proof follows. □

Proof of Proposition 1 : Let $\boldsymbol{\mu}_j$ and Σ_j be the location and the scale parameters of $f_j(\cdot)$. Define $R_j(\mathbf{x},F_j) = \{(\mathbf{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\}^{1/2}$. Under elliptic symmetry of $f_j(\cdot)$, we have $f_j(\mathbf{x}) = \Gamma(d/2)(2\pi)^{-d/2}|\Sigma_j|^{-1/2}g_j(R_j(\mathbf{x},F_j))/R_j(\mathbf{x},F_j)^{d-1}$, where $g_j(\cdot)$ is the p.d.f. of $R_j(\mathbf{x},F_j)$ (see e.g., Fang, Kotz and Ng 1989). Now, from Lemma 1, it follows that

$$d_B(\mathbf{x}) = \arg\max_j \pi_j f_j(\mathbf{x}) = \arg\max_j c_j\theta_j\{\mathrm{O}(\mathbf{x},F_j)\}/\{\mathrm{O}_j(\mathbf{x},F_j)\}^{d-1},$$

where $\theta_j(\cdot)$ is the density function of $\mathrm{O}(\mathbf{x},F_j)$. Since $\mathrm{PD}(\mathbf{x},F) = (1+\mathrm{O}(\mathbf{x},F))^{-1}$, the results of sampling distribution lead to Proposition 1. □

Proof of Theorem 2 : We split the proof into two lemmas (Lemma 2 & 3). Recall that, $r_{n_i}^{(i)}(\mathbf{x}_{jk}) = \hat{\rho}^*_{ih_i}(\hat{\delta}_{n_i}^{(i)}(\mathbf{x}_{jk}))(\hat{\delta}_{n_i}^{(i)}(\mathbf{x}_{jk}))^{d-3}/(1-\hat{\delta}_{n_i}^{(i)}(\mathbf{x}_{jk}))^{d-1}$ for $i,j = 1,2$ and $k = 1,2,\ldots,n_j$, where $\hat{\rho}^*$ stands for the usual kernel density estimate for $j \neq i$, and it stands for the leave one out density estimate for $j = i$. Also define $r^{(j)}(\mathbf{x}) = \rho_j(\delta^{(j)}(\mathbf{x}))(\delta^{(j)}(\mathbf{x}))^{d-3}/(1-\delta^{(j)}(\mathbf{x}))^{d-1}$ for $j = 1,2$.

Lemma 2 : Under the assumptions of Theorem 2, both for $i = 1$ and 2, $\sup_{\mathbf{X}} |r_{n_i}^{(i)}(\mathbf{x}) - r^{(i)}(\mathbf{x})| \xrightarrow{P} 0$ as $\min\{n_1, n_2\} \to \infty$.

Proof of Lemma 2 : Using the properties of kernel density estimate [41], under (A3) and uniform continuity of PD (follows from elliptic symmetry of $f_i(\cdot)$ [50]), we have, $\sup_{\delta} |\hat{\rho}_{ih_i}(\delta) - \rho_i(\delta)| \xrightarrow{P} 0$ as $n_i \to \infty$. Again, from the uniform convergence of PD [49], under elliptic symmetry of $f_i(\cdot)$, we have, $\sup_{\mathbf{X}} |\hat{\delta}_{n_i}^{(i)}(\mathbf{x}) - \delta^{(i)}(\mathbf{x})| \xrightarrow{P} 0$ as $n_i \to \infty$. This implies

$$\begin{aligned}\sup_{\mathbf{X}} |\hat{\rho}_{ih_i}(\hat{\delta}_{n_i}^{(i)}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))| &\leq \sup_{\mathbf{X}} |\hat{\rho}_{ih_i}(\hat{\delta}_{n_i}^{(i)}(\mathbf{x})) - \rho_i(\hat{\delta}_{n_i}^{(i)}(\mathbf{x}))| + \sup_{\mathbf{X}} |\rho_i(\hat{\delta}_{n_i}^{(i)}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))| \\ &\leq \sup_{y} |\hat{\rho}_{ih_i}(y) - \rho_i(y)| + \sup_{\mathbf{X}} |\rho_i(\hat{\delta}_{n_i}^{(i)}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))|.\end{aligned}$$

Using uniform continuity and vanishing at infinity properties of PD, for any given $\epsilon > 0$ we can find $M_{1\epsilon} > 0$ & $M_{2\epsilon} < 1$ such that $P(M_{1\epsilon} \leq \delta^{(i)}(\mathbf{x}) \leq M_{2\epsilon}) > 1 - \epsilon$. Now, if $\delta^{(i)}(\mathbf{x}) \in [M_{1\epsilon}, M_{2\epsilon}]$, it is easy to check that $\sup_{\mathbf{X}} |\hat{\rho}_{ih_i}(\hat{\delta}_{n_i}^{(i)}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))| \xrightarrow{a.s.} 0$ and $\sup_{\mathbf{X}} \left| \frac{(\hat{\delta}_{n_i}^{(i)}(\mathbf{X}))^{d-3}}{(1-\hat{\delta}_{n_i}^{(i)}(\mathbf{X}))^{d-1}} - \frac{(\delta^{(i)}(\mathbf{X}))^{d-3}}{(1-\delta^{(i)}(\mathbf{X}))^{d-1}} \right| \xrightarrow{a.s.} 0$ as $n_i \to \infty$. Lemma 2 follows from these results. □

Lemma 3 : Define

$$\Delta_n^{CV}(k) = \sum_{i=1, j\neq i}^{2} \frac{\pi_i}{n_i} \sum_{l=1}^{n_i} I\left\{ \frac{r_{n_j}^{(j)}(\mathbf{x}_{il})}{r_{n_i}^{(i)}(\mathbf{x}_{il})} \geq k_i \right\}, \; \Delta(k) = \sum_{i=1, j\neq i}^{2} \pi_i \, P\left\{ \frac{r^{(j)}(\mathbf{X})}{r^{(i)}(\mathbf{X})} \geq k_i \middle| \mathbf{X} \in i-th \; class \right\},$$

where $n{=}(n_1, n_2)$ and $k_i = 1/k$ for $i = 1$ and $k_i = k$ for $i = 2$. Also define, $c_n = \arg\min_k \; \Delta_n^{CV}(k)$ and $c = \arg\min_k \; \Delta(k)$. If $\Delta(\cdot)$ possess a unique minima, under the assumptions of Theorem 2, $c_n \xrightarrow{P} c$ as $\min\{n_1, n_2\} \to \infty$.

Proof of Lemma 3 : Since $\Delta(\cdot)$ has a unique minima, $\sup_k |\Delta_n^{CV}(k) - \Delta(k)| \xrightarrow{P} 0 \Rightarrow c_n \xrightarrow{P} c$. So we will prove that $\sup_k |\Delta_n^{CV}(k) - \Delta(k)| \xrightarrow{P} 0$ as $\min\{n_1, n_2\} \to \infty$. Note that

$$\begin{aligned}|\Delta_n^{CV}(k) - \Delta(k)| &\leq \sum_{i=1, j\neq i}^{2} \frac{\pi_i}{n_i} \sum_{l=1}^{n_i} \left| I\left\{ \frac{r_{n_j}^{(j)}(\mathbf{x}_{il})}{r_{n_i}^{(i)}(\mathbf{x}_{il})} \geq k_i \right\} - P\left\{ \frac{r^{(j)}(\mathbf{X})}{r^{(i)}(\mathbf{X})} \geq k_i \middle| \mathbf{X} \in i-th \; class \right\} \right| \\ &\leq \sum_{i=1, j\neq i}^{2} \frac{\pi_i}{n_i} \sum_{l=1}^{n_i} \left| I\left\{ \frac{r_{n_j}^{(j)}(\mathbf{x}_{il})}{r_{n_i}^{(i)}(\mathbf{x}_{il})} \geq k_i \right\} - I\left\{ \frac{r^{(j)}(\mathbf{x}_{il})}{r^{(i)}(\mathbf{x}_{il})} \geq k_i \right\} \right|\end{aligned}$$

$$+\sum_{i=1,j\neq i}^{2}\frac{\pi_i}{n_i}\sum_{l=1}^{n_i}\left|I\left\{\frac{r^{(j)}(\mathbf{x}_{il})}{r^{(i)}(\mathbf{x}_{il})}\geq k_i\right\}-P\left\{\frac{r^{(j)}(\mathbf{X})}{r^{(i)}(\mathbf{X})}\geq k_i\middle|\mathbf{X}\in i-th\ class\right\}\right|.$$

Recall that $\gamma(\mathbf{x})=r^{(2)}(\mathbf{x})/r^{(1)}(\mathbf{x})$, and let us consider the case $i=1$. Define

$$A_n(k_1)=\frac{1}{n_1}\sum_{i=1}^{n_1}\left|I\{\gamma(\mathbf{x}_{1i})\geq k_1\}-P\{\gamma(\mathbf{X})\geq k_1|\mathbf{X}\in 1st\ class\}\right| \text{ and}$$

$$B_n(k_1)=\frac{1}{n_1}\sum_{i=1}^{n_1}\left|I\{\hat{\gamma}_n(\mathbf{x}_{1i})\geq k_1\}-I\{\gamma(\mathbf{x}_{1i})\geq k_1\}\right|, \text{ where } \hat{\gamma}_n(\mathbf{x})=r^{(2)}_{n_2,h_2}(\mathbf{x})/r^{(1)}_{n_1,h_1}(\mathbf{x}).$$

Using the Glivenko-Cantelli theorem, one can show that $\sup_{k_1}|A_n(k_1)|\stackrel{a.s.}{\to}0$.

Under (A2), given any $\epsilon>0$ we get a $\delta_\epsilon>0$ such that $\sup_{k_1}|F_{\gamma,1}(k_1+\delta_\epsilon/2)-F_{\gamma,1}(k_1-\delta_\epsilon/2)|<\epsilon$. Using this δ_ϵ we get,

$$\begin{aligned}
B_n(k_1) &= \frac{1}{n_1}\sum_{i:|\gamma(\mathbf{x}_{1i})-k_1|\leq\delta_\epsilon/2}|I\{\hat{\gamma}_n(\mathbf{x}_{1i})<k_1\}-I\{\gamma(\mathbf{x}_{1i})<k_1\}| \\
&\quad+\frac{1}{n_1}\sum_{i:|\gamma(\mathbf{x}_{1i})-k_1|>\delta_\epsilon/2}|I\{\hat{\gamma}_n(\mathbf{x}_{1i})<k_1\}-I\{\gamma(\mathbf{x}_{1i})<k_1\}| \\
&\leq \frac{1}{n_1}\sum_{i=1}^{n_1}I(|\gamma(\mathbf{x}_{1i})-k_1|\leq\delta_\epsilon/2)+\frac{1}{n_1}\sum_{i:|\gamma(\mathbf{x}_{1i})-k_1|>\delta_\epsilon/2}|I\{\hat{\gamma}_n(\mathbf{x}_{1i})<k_1\}-I\{\gamma(\mathbf{x}_{1i})<k_1\}|
\end{aligned}$$

Note that $\frac{1}{n_1}\sum_{i=1}^{n_1}I(|\gamma(\mathbf{x}_{1i})-k_1|\leq\delta_\epsilon/2)\stackrel{a.s.}{\to}P(|\gamma(\mathbf{X}_1)-k_1|\leq\delta_\epsilon/2)<\epsilon$ [using (A2)] as $\min\{n_1,n_2\}\to\infty$. Now, using Lemma 2 and the relation between $f_j(\cdot)$ and $r^{(j)}(\cdot)$, $\hat{\gamma}_n(\mathbf{x})\stackrel{u}{\to}\gamma(\mathbf{x})$ in probability as $\min\{n_1,n_2\}\to\infty$. Hence, $|\gamma(\mathbf{x})-k_1|>\delta_\epsilon/2\Rightarrow\exists\ N_0\geq1$ such that for all $n=(n_1,n_2)$ with $\min\{n_1,n_2\}\geq N_0$, $|\hat{\gamma}_n(\mathbf{x})-k_1|>\delta_\epsilon/2$. Therefore, for all n with $\min\{n_1,n_2\}\geq N_0$, we have $\frac{1}{n_1}\sum_{i:|\gamma(\mathbf{x}_{1i})-k_1|>\delta_\epsilon/2}|I\{\hat{\gamma}_n(\mathbf{x}_{1i})<k_1\}-I\{\gamma(\mathbf{x}_{1i})<k_1\}|=0$ and hence $B_n(k_1)\leq\epsilon$. Now, using the same argument for $i=2$, we get the proof of Lemma 3.□

(Continuation of the proof of Theorem 2)

Note that $|\Delta_n-\Delta|\leq\sum_{j=1}^{2}\int\left|\prod_{i=1,i\neq j}^{2}I\left\{\frac{r^{(j)}_{n_j}(\mathbf{x})}{r^{(i)}_{n_j}(\mathbf{x})}\geq c_n\right\}-\prod_{i=1,i\neq j}^{2}I\left\{\frac{r^{(j)}(\mathbf{x})}{r^{(i)}(\mathbf{x})}\geq c\right\}\right|f_j(\mathbf{x})d\mathbf{x}$.

Using Lemma 2 & 3 and DCT (since indicators are bounded functions), we have $|\Delta_n-\Delta|\stackrel{P}{\to}0$.

Taking expectation w.r.t. the training sample and again using DCT, Theorem 2 is proved.□

Proof of Theorem 3: Note that if we can show that for any fixed $\mathbf{x}$, $p_n^*(1|\mathbf{x}) \xrightarrow{P} p(1|\mathbf{x})$ as $\min\{n_1, n_2\} \to \infty$, the rest of the proof follows from DCT. Now, we will prove it using the method of contradiction. If possible, let us assume that $p_n^*(1|\mathbf{x}) \stackrel{P}{\nrightarrow} p(1|\mathbf{x})$. So, $\exists\ \epsilon_0 > 0$ and a sub-sequence $\{n_k = (n_{1k}, n_{2k}) : k \geq 1\}$ such that $|p_{n_k}^*(1|\mathbf{x}) - p(1|\mathbf{x})| > \epsilon_0$ for all $k \geq 1$. Let $\{H_{n_k},\ k \geq 1\}$ be the corresponding sequence of bandwidth range. Since $p_{n_k}^*(1|\mathbf{x})$ is a weighted average of $\hat{p}_{n_k,h_1,h_2}(1|\mathbf{x})$ s, one can get a sub-sequence $\{(h_1^{n_k}, h_2^{n_k}) \in H_{n_k},\ k \geq 1\})$ such that $|\hat{p}_{n_k,h_1^{n_k},h_2^{n_k}}(1|\mathbf{x}) - p(1|\mathbf{x})| > \epsilon_0$ for all $k \geq 1$. So, along this sub-sequence $\hat{p}_{n_k,h_1^{n_k},h_2^{n_k}}(1|\mathbf{x}) \stackrel{P}{\nrightarrow} p(1|\mathbf{x})$. But this sequence of smoothing parameters satisfy the regularity condition (A3). So, we arrive at a contradiction. □

References

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

[2] Chakraborty, B. and Chaudhuri, P. (2003) On the use of genetic algorithm with elitism in robust and nonparametric multivariate analysis. *Aust. J. Statist.*, **32**, 13-27.

[3] 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.

[4] Christmann, A., Fischer, P. & 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.

[5] Cover, T. M. and Hart, P. E. (1967) Nearest neighbor pattern classification, *IEEE Trans. Info. Theory*, **13**, 21-27.

[6] Croux, C. & Dehon, C. (2001) Robust linear discriminant analysis using S-estimators. *Canad. J. Statist.*, **29**, 473-492.

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

[8] Fraiman, R., Liu, R.Y. & 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.

[9] Friedman, J. (1994) Flexible metric nearest neighbor classification. *Tech. Rep., Dept. of Stat., Stanford Univ.*

[10] Friedman, J. (1996) Another approach to polychotomous classification. *Tech. Rep., Dept. of Stat., Stanford Univ.*

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

[12] Ghosh, A. K. & Chaudhuri, P. (2005a) On data depth and distribution free discriminant analysis using separating surfaces. *Bernoulli*, **11**, 1-27.

[13] Ghosh, A. K. & Chaudhuri, P. (2005b) On maximum depth and related classifiers. *Scand. J. Statist.*, **32**, 328-350.

[14] Ghosh, A. K., Chaudhuri, P. and Murthy, C. A. (2005) On visualization and aggregation of nearest neighbor classifiers. *IEEE Trans. Pattern Anal. Machine Intell.,* **27**, 1592-1602.

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

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

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

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

[19] Hoberg, R. and Mosler, K. (2006) *DIMACS Series in Mathematics and Theoretical Computer Science*, (R. Liu and R. Serfling ed.), **72**, 49-59.

[20] Hodges, J. L. (1955) A bivariate sign test. *Ann. Math. Statist.*, **26**, 523-527.

[21] Holmes, C.C. and Adams, N.M. (2002) A probabilistic nearest-neighbour algorithm for statistical pattern recognition. *J. Roy. Statist. Soc. Ser. B.*, **64**, 295-306.

[22] Holmes, C.C. and Adams, N.M. (2003) Likelihood inference in nearest-neighbour classification models. *Biometrika*, **90** , 99-112.

[23] Hubert, M. & Van Driessen, K. (2004) Fast and robust discriminant analysis. *Comput. Statist. Data Anal.*, **45**, 301-320.

[24] Jornsten, R. (2004) Clustering and classification based on the L1 data depth. *J. Multivariate Anal.*, **90**, 67-89.

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

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

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

[28] López-Pintado, S. and Romo, J. (2006) *DIMACS Series in Mathematics and Theoretical Computer Science*, (R. Liu and R. Serfling ed.), **72**, 103-119.

[29] Mahalanobis, P. C. (1936) On the generalized distance in statistics. *Proc. Nat. Acad. Sci., India*, **12**, 49-55.

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

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

[32] Oja, H. and Randles, R. (2004) Multivariate nonparametric tests. *Statist. Science*, **19**, 598-605.

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

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

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

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

[37] Rousseeuw, P. J. and Ruts, I. (1996) Algorithm AS 307: bivariate location depth. *Appl. Statist. (JRSS-C)*, **45**, 516-526.

[38] Schapire, R.E., Fruend, Y., Bartlett, P., and Lee, W. (1998). Boosting the Margin: A New Explanation for the Effectiveness of Voting Method. *Ann. Statist.*, **26**, 1651-1686.

[39] Seber, G. A. F. and Wild, C. J. (1989) *Nonlinear Regression*, Wiley, New York.

[40] 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.

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

[42] Singh, K. (1991) A notion of majority depth. *Tech. Rep., Dept. of Stat., Rutgers Univ.*.

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

[44] Tyler, D. E. (1987) A distribution free M-estimator of multivariate scatter. *Ann. Statist.,* **15**, 234-251.

[45] Vardi, Y. & Zhang, C. H. (2000) The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, **97**, 1423-1426.

[46] Wilcox, R. R. (2005) *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press.

[47] Xia, C., Lin, L. and Yang, G. (2008) An Extended Projection Data Depth and Its Applications to Discrimination . *Comm. Statist. - Theory & Methods*, **37**, 2276-2290.

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

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

[50] Zuo, Y. (2003) Projection based depth functions and associated medians. *Ann. Statist.*, **31**, 1460-1490.