Ann Inst Stat Math (2012) 64:657–676
DOI 10.1007/s10463-011-0324-y

On robust classification using projection depth

Subhajit Dutta · Anil K. Ghosh

Received: 11 August 2009 / Revised: 29 November 2010 / Published online: 31 March 2011

Abstract This article uses projection depth (PD) for robust classification of multivariate data. Here we consider two types of classifiers, namely, the maximum depth classifier and the modified depth-based classifier. The latter involves kernel density estimation, where one needs to choose the associated scale of smoothing. We consider both the single scale and the multi-scale versions of kernel density estimation, and investigate the large sample properties of the resulting classifiers under appropriate regularity conditions. Some simulated and real data sets are analyzed to evaluate the finite sample performance of these classification tools.

Keywords Bayes risk · Bandwidth · Cross-validation · Data depth · Elliptic symmetry · Kernel density estimation · Misclassification rate · Multi-scale smoothing

1 Introduction

Over the last three decades, data depth has emerged as a powerful tool for statistical analysis of multivariate data. Robust estimation of multivariate location and scatter (see, e.g., Liu et al. 1999), test of statistical hypothesis (see, e.g., Chaudhuri and Sengupta 1993; Liu and Singh 1993), detection of outliers (see, e.g., Chen et al. 2009), supervised and unsupervised classification (see, e.g., Hoberg 2000; Jornsten 2004; Ghosh and Chaudhuri 2005a,b) are some examples of its wide spread applications.

S. Dutta (✉) · A. K. Ghosh
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203, B. T. Road, Kolkata 700108, India
e-mail: subhajit_r@isical.ac.in

A. K. Ghosh
e-mail: akghosh@isical.ac.in

Ghosh and Chaudhuri (2005b) introduced the maximum depth classifier and also proposed a modified version of it based on half-space depth (HD) (see, e.g., Tukey 1975). But, as pointed out in Li et al. (2011), estimation of too many parameters makes this modification quite complicated and not of much practical use. Zuo (2003, 2004) also pointed out some drawbacks of HD, especially in the context of breakdown points. These necessitate the development of better depth-based classification tools, which have good theoretical properties and are computationally tractable and hence useful in practice. Here, we use projection depth (see, e.g., Zuo and Serfling 2000a) for this purpose. PD satisfies all four desirable properties of depth (i.e., affine invariance, maximality at center, monotonicity w.r.t. the deepest point and vanishing at infinity), and it has several advantages over many existing depth functions (see Sect. 2). PD of an observation $\mathbf{x}$ w.r.t. a multivariate distribution F is defined as $\mathrm{PD}(\mathbf{x}, F) = \{1 + \mathrm{O}(\mathbf{x}, F)\}^{-1}$, where $\mathrm{O}(\mathbf{x}, F) = \sup_{\alpha:\|\alpha\|=1}\{|\boldsymbol{\alpha}'\mathbf{x} - m_F(\boldsymbol{\alpha}'\mathbf{X})|/\sigma_F(\boldsymbol{\alpha}'\mathbf{X})\}$, for m_F and σ_F being some univariate measures of location and scatter, respectively. Zuo and Serfling (2000a) used median and median absolute deviation (MAD) about median as these measures, but one can use other measures as well. The empirical version of PD is obtained by replacing F by its empirical analog F_n, which puts a mass of $1/n$ on each of the n data points. Under some regularity conditions, the contours of this empirical depth function converge to their population analogs (see, e.g., He and Wang 1997; Zuo and Serfling 2000b). In the next section, we will use this empirical version for maximum depth classification, and the resulting classifier will be referred to as the maximum PD classifier.

2 Maximum projection depth classifier

If an observation is a proper representative of a class, it is expected to have higher depth with respect to that class. Based on this simple idea, a maximum depth classifier classifies an observation to the class for which it has the maximum depth. In principle, any depth function can be used for maximum depth classification, but in high dimensions, due to computational difficulty, it is not feasible to use some of them. Mahalanobis depth (MD) (see Liu et al. 1999) 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 can plug in robust estimates for the location and the scatter (see, e.g., Tyler 1987; Rousseeuw and Van Driessen 1999; Croux and Dehon 2001), but the computational simplicity gets lost. L_1 depth (L_1D) (see, e.g., Vardi and Zhang 2000) is also easy to compute, but its usual version is not affine invariant. An affine invariant version is given by $L_1D(\mathbf{x}, F) = 1 - \|E_F\left\{\boldsymbol{\Sigma}_F^{-1/2}(\mathbf{x} - \mathbf{X})/\|\boldsymbol{\Sigma}_F^{-1/2}(\mathbf{x} - \mathbf{X})\|\right\}\|$, and here also, one needs to plug in a robust estimate for the scatter matrix $\boldsymbol{\Sigma}_F$ to get its robust empirical analog. Though HD satisfies all four desirable properties of a depth function and can be computed in high dimensions (see, e.g., Rousseeuw and Struyf 1998; Ghosh and Chaudhuri 2005a), a major limitation of HD is the stepwise nature of its empirical version. As a result, an observation can have maximum HD with respect to multiple classes. This problem is more serious when a test observation lies outside the convex hull formed by the training data from different classes. In that case, it has zero depth with respect to all competing classes, and the maximum HD classifier often misclassifies it. We also have similar problems for other depth functions like simplicial depth (SD), majority depth and convex hull peeling depth (see, e.g., Liu et al. 1999), which are based on counting. Unlike these depth functions, both the empirical and the population versions of PD are continuous in $\mathbf{x}$ (see, e.g., Zuo (2003) for uniform continuity of PD under mild conditions) and are always positive. Hence, one does not have to deal with ties and observations with zero depth. Moreover, better breakdown properties of PD (see, e.g., Zuo 2003, 2004) are expected to make the resulting classifier robust against outliers.

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

$$d_1(\mathbf{x}) = \arg \max_{j \in \{1,\ldots,J\}} \mathrm{PD}(\mathbf{x}, F_{jn_j}) = \arg \min_{j \in \{1,\ldots,J\}} \mathrm{O}(\mathbf{x}, F_{jn_j}),$$

where F_{jn_j} is the empirical distribution function of the jth class ($j = 1, \ldots, J$). If the population density function f_j is unimodal and elliptically symmetric (see, e.g., Fang et al. 1989), $\mathrm{PD}(\mathbf{x}, F_j)$ turns out to be a monotonically increasing function of $f_j(\mathbf{x})$. So, when the f_js differ only in their location, and the priors of the competing classes are equal (i.e., $\pi_1 = \cdots = \pi_J = 1/J$), the population version of $d_1(\cdot)$ coincides with the Bayes classifier $d_B(\mathbf{x}) = \arg\max_{1 \le j \le J} \pi_j f_j(\mathbf{x})$. In such cases, the error rate of the classifier $d_1(\cdot)$ also converges to the Bayes risk Δ_B [the error rate of $d_B(\cdot)$] as the training sample size increases.

Theorem 1 *If $f_1, f_2, \ldots, f_J$ are elliptically symmetric, unimodal and they satisfy a location shift model (i.e., $f_j(\mathbf{x}) = f(\mathbf{x} - \boldsymbol{\mu}_j)$ for some common density function f and location parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_J$), the misclassification rate of the maximum PD classifier $d_1(\cdot)$ converges to the Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.*

2.1 Performance on simulated data sets

We analyze some simulated data sets to compare the performance of the maximum PD classifier with some other maximum depth classifiers. Here, we restrict ourselves to two class problems in two dimensions so that classifiers based on computationally expensive depth functions like SD can also be used for comparison. To compute the empirical PD of an observation in $\mathbb{R}^2$, we search for $\boldsymbol{\alpha} = (\sin\theta, \cos\theta)$ over a fine grid (one can also use the R package ExPD2D for computing PD for observations in $\mathbb{R}^2$). As we have mentioned earlier, any measures of univariate location and scale can be used to define PD. Here, we used median as the measure of location, and both MAD and quartile deviation (QD) as the measure of scale. Since there was no visible difference in the performance of the resulting classifiers, here we have reported the results based on QD only. Note that if we use MD for maximum depth classification, it leads to linear discriminant analysis (LDA). Here we have reported the error rates of LDA and its robust versions, where MCD estimates (see, e.g., Rousseeuw and Van Driessen 1999) of the location and the scatter are plugged in. Here, we have used two MCD estimates, one based on 50% observations (which has the highest breakdown) and the other based on 75% observations [suggested in Hubert and Van Driessen (2004) for good finite sample efficiency]. The resulting robust classifiers will be referred to as the $\text{MD}_{1/2}$ classifier and the $\text{MD}_{3/4}$ classifier, respectively. For the affine invariant, robust version of L_1D, we used the MCD estimate with the highest breakdown. Throughout this section, we consider the prior probabilities of the competing classes to be equal.

We begin with an example involving two normal distributions having the same dispersion matrix $\mathbf{I_2}$, the 2×2 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 200 and a test set of size 500 taking equal number of observations from two competing classes. This procedure is repeated 500 times, and the average test set error rates of different maximum depth classifiers and their corresponding standard errors (SE) are reported in Table 1. Bayes risks are also reported to facilitate comparison. Note that for HD and SD, the observations having zero depth with respect to all competing classes were classified using the nearest neighbor (see, e.g., Cover and Hart 1967) algorithm. Otherwise, the error rates of these two methods would have been much higher. As expected, LDA yielded the best performance in these examples, and its error rates were close to the corresponding Bayes risks. The performance of its robust versions ($\text{MD}_{1/2}$ and $\text{MD}_{3/4}$), the L_1D classifier and the PD classifier was also competitive. However, when we repeated the same experiment using Cauchy distributions (which is heavy-tailed), in the presence of outlying observations, LDA failed to perform well (see Table 1). In these cases, the performance of all other depth-based classifiers was significantly better than LDA, which shows the robustness of these depth-based methods.

In the above examples, the overall performance of the robust MD classifiers was better than the PD classifier. This is probably due to the fact that in MD and robust MD, we used the information about the homoscedastic structure of two competing populations, but for PD, we could not use this fact. As a result, like the Bayes classifier, MD and robust MD classifiers were always linear, but the PD classifier led to a nonlinear estimate of the class boundary in some cases. However, this nature can be helpful in some situations. To demonstrate this, we consider an example with two normal distributions having location parameters (0, 0) and (2, 2), and the common dispersion matrix $\mathbf{I_2}$. Here we replace 10% of the class-1 observations by observations from a normal distribution with the mean (20, 20) and the scatter matrix $\mathbf{I_2}$. A scatter plot of this data set is given in Fig. 1, where the 'dots' $(\cdot)$ and the 'crosses' $(\times)$ represent the observations from the two classes. In the presence of these contaminating observations, LDA was most affected, and it misclassified almost half of the observations

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

	μ	Bayes risk	LDA	HD	SD	L_1D	$\text{MD}_{1/2}$	$\text{MD}_{3/4}$	PD
Normal	1	23.98	24.07(0.08)	24.96(0.09)	25.07(0.09)	24.40(0.09)	24.30(0.09)	24.17(0.08)	24.83(0.09)
	2	7.87	7.99(0.05)	8.38(0.06)	8.44(0.06)	8.14(0.06)	8.12(0.06)	8.06(0.05)	8.33(0.06)
Cauchy	1	30.40	43.57(0.41)	34.03(0.15)	34.01(0.14)	33.17(0.13)	30.85(0.09)	30.90(0.13)	32.31(0.10)
	2	19.58	33.85(0.60)	22.86(0.13)	23.10(0.12)	22.04(0.11)	19.93(0.08)	19.96(0.08)	20.82(0.09)

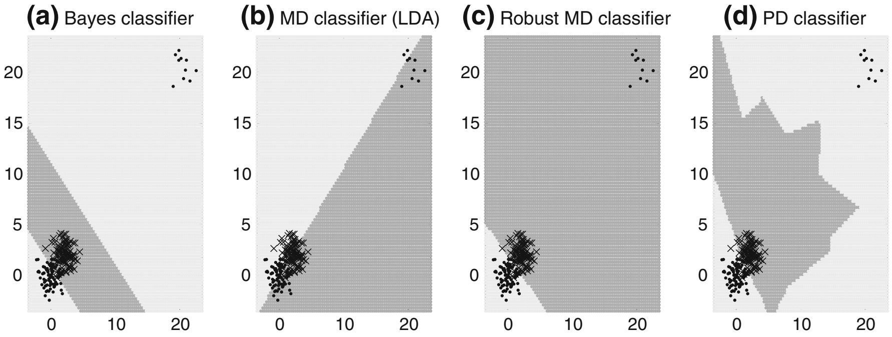

Fig. 1 Class boundaries estimated by different maximum depth classifiers

(see Fig. 1 for class boundaries estimated by different methods, where the light and the dark regions indicate decisions in favor of the first and the second class, respectively). We repeated this experiment 500 times, but in most of the cases, LDA showed the same behavior. However, the robust MD classifiers and the PD classifier were less affected (see Fig. 1c and d). Over these 500 trials, $\mathrm{MD}_{1/2}$ and $\mathrm{MD}_{3/4}$ classifiers had average error rates of 12.71 and 12.65%, respectively, with the same SE 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). But, alike the Bayes classifier (see Fig. 1a), the PD classifier correctly classified these observations (see Fig. 1d). As a result, it had a significantly lower error rate of 10.78% with an SE of 0.09%. For L_1D, HD and SD, these error rates were 12.68, 22.40 and 24.56%, respectively, with the corresponding SE of 0.06, 0.14 and 0.16%. When 10% outlying observations were not considered for computing misclassification rates, average test set error rates of LDA turned out to be 42.63% with an SE of 0.22%. Maximum depth classifiers based on HD (error rate 22.51%, SE 0.14%) and SD (error rate 24.65%, SE 0.16%) also had higher misclassification rates compared to $\mathrm{MD}_{1/2}$, $\mathrm{MD}_{3/4}$, L_1D and PD classifiers. For these four methods, average error rates were 8.48, 8.41, 9.03 and 8.98%, respectively, with a common SE of 0.06%. This shows better robustness properties of these four classifiers.

3 Modified projection depth classifier

The maximum PD classifier and other maximum depth classifiers described in Sect. 2 perform well when the priors are equal, and the population distributions differ only in their location. However, in practice, different populations may have different priors, and the population distributions may also differ in their scatters and shapes. In such situations, the maximum depth classifiers may not perform well, and they need to be modified. Xia et al. (2008) proposed a modification of the PD classifier that can be used when the populations have different scatters. Li et al. (2011) also proposed some modifications assuming a monotonic relationship between the depth and the density functions. In this section, we propose a modification which works under a more general set-up. This modification is also motivated by elliptic symmetry of the underlying distributions. Note that if the population distributions are elliptic, the Bayes classifier is given by $d_B(\mathbf{x}) = \arg\max_{1\le j\le J} \pi_j \psi_j\{D(\mathbf{x}, F_j)\}$, where $D(\mathbf{x}, F_j)$ is the depth of $\mathbf{x}$ w.r.t F_j and ψ_j is an appropriate transformation function (see Ghosh and Chaudhuri 2005b). If the population distributions differ only in their location, the ψ_js are same for all populations. Further, if the f_js are unimodal, then ψ_js are monotonically decreasing. In such cases, if the π_js are equal, the Bayes classifier turns out to be the maximum depth classifier. But, when even one of these assumptions fails to hold, one needs to know functional forms of the ψ_js. For PD, ψ_j can be easily obtained (see Lemma 1 in Appendix), and using that one arrives at the following proposition.

Proposition 1 *If $f_1, f_2, \ldots, f_J$ are elliptically symmetric, the Bayes classifier is given by*

$$d_{\mathrm{B}}(\mathbf{x}) = \arg\max_{j\in\{1,\ldots,J\}} \lambda_j \rho_j\{PD(\mathbf{x}, F_j)\}\{PD(\mathbf{x}, F_j)\}^{d-3}/\{1 - PD(\mathbf{x}, F_j)\}^{d-1},$$

where $\rho_j(\cdot)$ is the density function of $PD(\mathbf{x}, F_j)$, and λ_j is an appropriate constant.

Proposition 1 holds for any definition of PD, while the λ_js change depending on the choice of univariate measures of location and scale (see the proof of Lemma 1 for a clear idea). To construct the modified PD classifier, we estimate PD($\mathbf{x}$, F_j) by its sample analog PD($\mathbf{x}$, F_{jn_j}), and ρ_j is estimated using the kernel density estimation technique (see, e.g., Silverman 1986). Note that irrespective of the dimension of the measurement space, here we need only one-dimensional density estimation. This helps us to get rid of the curse of dimensionality that one usually faces in high-dimensional nonparametric density estimation. For estimation of the ρ_j $(1 \le j \le J)$, one has to choose the bandwidth h_j as well. For a given value of h_j, this density estimate is given by $\hat{\rho}_{jh_j}(\delta) = (n_j h_j)^{-1} \sum_{i=1}^{n_j} K\left\{h_j^{-1}(\delta - \hat{\delta}_{n_j}^{(j)}(\mathbf{x}_{ji}))\right\}$, where $\hat{\delta}_{n_j}^{(j)}(\mathbf{x}) =$ PD($\mathbf{x}$, F_{jn_j}), and K is the kernel function. Throughout this article, we will assume that K has bounded first derivative, and for all numerical studies, we will use the Gaussian kernel, which satisfies this property. In a two-class problem, the resulting classifier can be expressed as

$$d_2(\mathbf{x}) = \begin{cases} 1 & \text{if } \log[r^{(1)}_{n_1,h_1}(\mathbf{x})] - \log[r^{(2)}_{n_2,h_2}(\mathbf{x})] > k, \\ 2 & \text{otherwise,} \end{cases}$$

where $r^{(j)}_{n_j,h_j}(\mathbf{x}) = \hat{\rho}_{jh_j}(\hat{\delta}^{(j)}_{n_j}(\mathbf{x}))(\hat{\delta}^{(j)}_{n_j}(\mathbf{x}))^{d-3}/(1 - \hat{\delta}^{(j)}_{n_j}(\mathbf{x}))^{d-1}$ for $j = 1, 2$, and $k = \log(\lambda_2/\lambda_1)$. Clearly, the performance of the classifier $d_2(\cdot)$ depends on the choice of h_1, h_2 and k. If h_1 and h_2 satisfy the assumption (A3), and k is chosen by minimizing the cross-validation estimate of the error rate, under the assumptions (A1) and (A2) [(A1)–(A3) are mentioned in the statement of Theorem 2], the error rate of the modified PD classifier $d_2(\cdot)$ converges to the Bayes risk as the training sample size increases.

Theorem 2 *Suppose that f_1 and f_2 are elliptically symmetric. Also consider the following assumptions*

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

(A2) *For* $j = 1, 2$, $F_{\gamma,j}(z) = P(\gamma(\mathbf{X}) \le z \mid \mathbf{X} \in j$*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* $\delta^{(j)}(\mathbf{x}) = \mathrm{PD}(\mathbf{x}, F_j)$.

(A3) *For* $j = 1, 2$, $h_j \to 0$ *and* $n_j h_j^4 \to \infty$ *as* $n_j \to \infty$.

If (A1)–(A3) *hold, and k is chosen by minimizing the cross-validation estimate of the error rate, the misclassification rate of the modified PD classifier* $d_2(\cdot)$ *converges to the Bayes risk as* $min\{n_1, n_2\} \to \infty$.

A similar classifier based on HD was proposed in Ghosh and Chaudhuri (2005b), which is quite complicated and has problems with ties and zero depths (as mentioned earlier). The empirical version of HD takes only discrete values, and this leads to a loss of information for continuous distributions. As a result, we often have poor density estimates with peaks near those discrete values. Moreover, because of the presence of many observations with zero depth, the resulting density estimate $\hat{f}_j$ has bumps in the tail, which is not desirable. On the contrary, the empirical version of PD is continuous in $\mathbf{x}$, and it does not have such problems. As a result, the modified PD classifier often outperforms the modified HD classifier.

Theorem 2 gives an idea about the optimal asymptotic order of the h_js. But, in practice, we need to estimate them from the data. Here, we used the leave-one-out cross-validation method to choose h_1 and h_2 along with k. Since the bandwidths are supposed to be proportional to the population dispersions, to reduce the computing cost, we chose $h_1 = (s_1/s_2)h_2$, where s_j $(j = 1, 2)$ is a dispersion measure (for robustness, here we used sample quartile deviation) of the estimated depth functions $\left\{\hat{\delta}^{(j)}_{n_j}(\mathbf{x}_{j1}), \hat{\delta}^{(j)}_{n_j}(\mathbf{x}_{j2}), \ldots, \hat{\delta}^{(j)}_{n_j}(\mathbf{x}_{jn_j})\right\}$. For a given h_2 (and $h_1 = (s_1/s_2)h_2$), we computed $r^{(i)}_{n_i,h_i}(\mathbf{x}_{jl}) = \hat{\rho}^*_{ih_i}(\hat{\delta}^{(i)}_{n_i}(\mathbf{x}_{jl}))(\hat{\delta}^{(i)}_{n_i}(\mathbf{x}_{jl}))^{d-3}/(1 - \hat{\delta}^{(i)}_{n_i}(\mathbf{x}_{jl}))^{d-1}$ for $i, j = 1, 2$ and $l = 1, \ldots, n_j$, where $\hat{\rho}^*$ stands for the leave-one-out (usual) kernel density estimate for $j = i$ $(j \ne i)$. The constant k was 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, and chose the one that led to the lowest cross-validation error rate. Due to stepwise nature of the cross-validation error rate, often we have multiple minimizers. In such cases, following Ghosh and Chaudhuri (2004) we chose the maximum of the optimizers.

A similar generalization is also possible for depth-based classification using MD and robust MD. Under appropriate conditions, the MCD estimate of the scatter matrix $\boldsymbol{\Sigma}_F$ converges to $c_F \boldsymbol{\Sigma}_F$, where c_F is a scalar that depends on the underlying distribution F (see, e.g., Cator and Lopuhaä 2011). However, 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 be adopted to develop a modified robust MD classifier, and its asymptotic optimality can be proved following the proof of Theorem 2.

For classification among $J(> 2)$ classes, $h_1, h_2, \ldots, h_J$ and $\lambda_1, \lambda_2, \ldots, \lambda_J$ can be chosen in a similar way, but it is computationally difficult to minimize the cross-validation error rate w.r.t. several parameters. Therefore, we consider a pair of classes at a time and perform $\binom{J}{2}$ binary classifications as discussed above. The results of all pairwise classifications are combined using the method of majority voting, where ties are resolved arbitrarily. Following the arguments of the proof of Theorem 2, under similar regularity conditions, one can prove the Bayes risk consistency of the modified PD classifier for these multi-class problems.

3.1 Performance on simulated data sets

To show the utility of the proposed modification, we begin with the same examples with the normal and the Cauchy distributions considered in Sect. 2, but here we consider three different choices for π_1 (0.5, 0.6 and 0.7). Note that for $\pi_1 = 0.5$, results of different maximum depth classifiers were reported in Table 1. As compared to them, the modified depth-based classifiers had marginally higher error rates in some cases (see Table 2). It is expected here since we do not use any information about the location shift model, and in addition to depths, we need to estimate the density functions. But, irrespective of prior probabilities, when the maximum depth classifiers led to the same error rate, the modified depth-based classifiers could reduce it significantly when $\pi_1 \neq \pi_2$. For $\pi_1 = 0.7$ and even for $\pi_1 = 0.6$, the modified PD classifier yielded error rates much lower than that of the maximum PD classifier. We observed the same phenomenon for the modified MD and the modified HD classifiers as well. One should notice that in almost all these examples, especially in the case of Cauchy distributions, modified versions of the PD and the robust MD classifiers significantly outperformed the modified HD classifier proposed in Ghosh and Chaudhuri (2005b).

Table 2 Misclassification rates (in %) of modified 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	23.98	23.10	20.41	7.86	7.66	7.00
	LDA	24.07 (0.08)	23.21 (0.08)	20.55 (0.08)	7.99 (0.05)	7.73 (0.05)	7.07 (0.05)
	HD	25.69 (0.10)	24.98 (0.10)	22.61 (0.10)	8.52 (0.06)	8.37 (0.06)	7.90 (0.06)
	MD	24.99 (0.09)	24.23 (0.09)	21.80 (0.09)	8.40 (0.06)	8.17 (0.06)	7.57 (0.06)
	$\mathrm{MD}_{1/2}$	25.32 (0.10)	24.52 (0.09)	22.17 (0.09)	8.58 (0.06)	8.31 (0.06)	7.66 (0.06)
	$\mathrm{MD}_{3/4}$	25.09 (0.10)	24.39 (0.09)	22.00 (0.09)	8.51 (0.06)	8.26 (0.06)	7.61 (0.06)
	PD	25.40 (0.10)	24.59 (0.10)	21.87 (0.09)	8.59 (0.06)	8.39 (0.06)	7.76 (0.06)
Cauchy distribution	Bayes risk	30.40	28.88	25.01	19.60	18.77	16.68
	LDA	43.57 (0.41)	40.25 (0.03)	30.32 (0.02)	33.85 (0.60)	39.89 (0.07)	30.39 (0.03)
	HD	34.49 (0.14)	33.28 (0.14)	29.32 (0.13)	23.12 (0.12)	22.54 (0.12)	20.55 (0.11)
	MD	38.88 (0.25)	36.72 (0.19)	30.70 (0.10)	26.75 (0.27)	26.09 (0.26)	23.92 (0.21)
	$\mathrm{MD}_{1/2}$	32.28 (0.10)	31.17 (0.11)	27.71 (0.13)	20.95 (0.09)	20.50 (0.09)	18.70 (0.09)
	$\mathrm{MD}_{3/4}$	32.43 (0.11)	31.21 (0.11)	27.77 (0.13)	21.00 (0.09)	20.54 (0.09)	18.58 (0.09)
	PD	32.26 (0.10)	31.25 (0.10)	27.66 (0.11)	20.94 (0.09)	20.45 (0.09)	18.86 (0.10)

Now consider some examples, where $\pi_1 = \pi_2$, and the two competing classes have the same location parameter (0, 0), but different scatter matrices $\mathbf{I_2}$ and $\sigma^2\mathbf{I_2}$, respectively, for the first and the second classes. Here also, we considered examples with the normal and the Cauchy distributions, and computed the error rates of the modified depth-based classifiers for two different values of σ^2 (4 and 9). Since the optimal class boundary is quadratic in these problems, to facilitate comparison, the error rates of quadratic discriminant analysis (QDA) and the Bayes risks are also reported in Table 3. Throughout this section, we used training and test samples of size 200 and 500, respectively, and the results are reported based on 500 trials. In all these examples, maximum depth classifiers led to almost 50% error rates, but their modified versions worked significantly better. In the case of normal distributions, as expected, QDA led to the best performance, but the overall performance of the modified PD classifier was significantly better than other depth-based classifiers considered here. In the case of Cauchy distributions, PD significantly outperformed QDA, the modified MD and the modified HD classifiers. We also analyzed two other data sets, one with normal and the other with Cauchy distributions, where the population distributions differ both in location [(0, 0) and (2, 2)] and scatter ($\mathbf{I_2}$ and $9\mathbf{I_2}$). Superiority of the modified PD classifier was evident even in these two cases.

Next, we consider some cases, where the location and the scatter parameters of the two populations are same, but they differ in their shapes. Again, the maximum depth classifiers yielded error rates close to 50%, but their modified versions showed considerable improvement. Along with the error rates of these modified versions, the error rates of QDA and the Bayes risks are reported in Table 4. We start with a classification problem between a standard bivariate normal distribution and a standard bivariate Cauchy distribution. QDA had the lowest error rate in this example, but those of the modified depth-based classifiers were competitive. Modified versions of the MD and the PD classifiers performed better than the modified HD classifier. Next, we consider a classification problem with two standard bivariate normal distributions, where one of them is truncated to have $\mathbf{x}$ with $\|\mathbf{x}\| \geq 4$. In this case, QDA and the modified depth-based classifiers, except $\text{MD}_{1/2}$, performed quite well, with MD and HD having an edge. But, when we carried out the same experiment with Cauchy distributions, QDA (which is based on non-robust estimates) had error rates close to

Table 3 Misclassification rates (in %) of modified depth-based classifiers and their standard errors

	Normal			Cauchy		
	$\mu = 0, \sigma^2 = 4$	$\mu = 0, \sigma^2 = 9$	$\mu = 1, \sigma^2 = 9$	$\mu = 0, \sigma^2 = 4$	$\mu = 0, \sigma^2 = 9$	$\mu = 1, \sigma^2 = 9$
Bayes risk	26.37	16.22	14.98	37.00	30.15	27.61
QDA	26.84 (0.08)	16.51 (0.07)	15.27 (0.07)	48.25 (0.13)	46.97 (0.27)	45.80 (0.19)
HD	31.88 (0.19)	24.94 (0.25)	23.14 (0.23)	41.28 (0.14)	34.37 (0.15)	32.12 (0.15)
MD	29.22 (0.11)	18.21 (0.09)	17.51 (0.10)	42.32 (0.14)	35.57 (0.17)	33.62 (0.15)
$\text{MD}_{1/2}$	30.16 (0.12)	19.04 (0.09)	17.90 (0.10)	41.32 (0.14)	33.78 (0.11)	30.89 (0.11)
$\text{MD}_{3/4}$	29.43 (0.12)	18.45 (0.09)	17.68 (0.10)	41.37 (0.13)	33.77 (0.12)	30.77 (0.12)
PD	28.86 (0.11)	17.89 (0.09)	16.64 (0.10)	40.44 (0.13)	32.63 (0.12)	30.05 (0.12)

Table 4 Misclassification rates (in %) of modified depth-based classifiers and their standard errors

	Bayes risk	QDA	HD	MD	$MD_{1/2}$	$MD_{3/4}$	PD
Normal versus Cauchy	33.57	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 versus Trun. normal	6.77	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 versus Trun. Cauchy	22.36	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 versus Mix. normal	21.75	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 versus Mix. Cauchy	38.28	50.00 (0.19)	43.95 (0.15)	45.19 (0.14)	45.48 (0.15)	45.69 (0.15)	41.99 (0.13)

50%, while the modified depth-based classifiers had significantly lower error rates. We also consider an example, where one class is bivariate normal with the mean (0, 0) and the dispersion matrix $25\mathbf{I_2}$, and the other one is an equal mixture of two bivariate normal distributions having the same mean (0, 0) but different dispersion matrices $\mathbf{I_2}$ and $100\mathbf{I_2}$. In this example, while all other classifiers had average error rates higher than 30%, the modified PD classifier yielded an average error rate (24.74 %) close to the Bayes risk. We observed the same phenomenon when the experiment was carried out with Cauchy distributions.

To study the robustness of these modified depth-based classifiers, we again consider the classification problem between $\mathrm{N}_2(\mathbf{0}, \mathbf{I}_2)$ and $\mathrm{N}_2(\mathbf{0}, 9\mathbf{I}_2)$ distributions. Figure 2 shows the optimum class boundary for this problem. We generated observations from two competing classes as before, but 10% of the training set observations of class-1 were replaced by outliers generated from a normal distribution with the location parameter (10, 10) and the scatter matrix $\mathbf{I_2}$ (see the scatter plot of the data set in Fig. 2a). In the presence of these outlying observations, QDA and modified MD classifier could not properly estimate the class boundary, but those for the modified robust MD classifier and the modified PD classifier were close to the optimum one (see

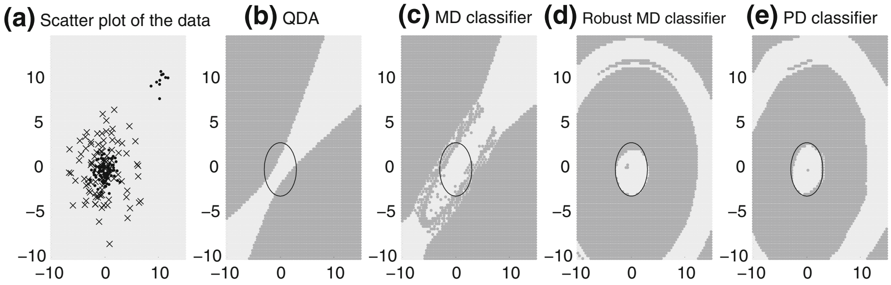

Fig. 2 Scatter plot of the data set and class boundaries estimated by different modified depth-based classifiers

Fig. 2). To compare the error rates of different classifier, we generated 500 training and test sets of size 200 and 500, but unlike the training sample, no outlier is added to the test set. In this problem, QDA and the modified MD classifier had average error rates of 30.45% (SE 0.11%) and 29.25% (SE 0.11%), respectively, but $\text{MD}_{1/2}$ (error rate 19.09%, SE 0.11%) and $\text{MD}_{3/4}$ (error rate 18.67%, SE 0.10%) classifiers performed much better. The modified PD classifier yielded the best performance in this example. It had an average error rate of 18.32% with an SE of 0.10%. We also considered a variant of this example, where 10% outliers were generated from an equal mixture of four normal distributions with location parameters (10, 10), (−10, 10), (10, −10) and (−10, −10) and the common scatter matrix $\mathbf{I_2}$. Even in this example, the robust MD classifiers and the PD classifier showed better robustness properties. While QDA led to an error of nearly 50%, those for $\text{MD}_{1/2}$ and $\text{MD}_{3/4}$ and PD classifiers were 19.07, 18.67 and 17.99% with the corresponding SE of 0.11, 0.11 and 0.09%, respectively. The modified MD classifier had an error rate of 20.82% with an SE of 0.13%. Next we consider a case, where two normal distributions differ both in the location ((0, 0) and (2, 2)) and scatter ($\mathbf{I_2}$ and $9\mathbf{I_2}$). Here also 10% of the class-1 observations in the training set were replaced by observations from a normal distribution with the location parameter (10, 10) and the scatter matrix $\mathbf{I_2}$. Similarly, 10% of the class-2 observations were also replaced by observations from a normal distribution with the location parameter (−10, −10) and the scatter matrix $\mathbf{I_2}$. In the presence of these two sets of outliers in the training sample, QDA again misclassified almost half of the test set observations. The modified MD classifier also had a higher error rate of 30.25% with an SE of 0.12%. But modified robust MD classifiers, $\text{MD}_{1/2}$ and $\text{MD}_{3/4}$, could reduce this error rate to 18.10% (SE 0.10%) and 17.87% (SE 0.10%), respectively. The modified PD classifier performed even better. It yielded an average error rate of 16.94% with an SE of 0.10%. These examples with clusters of outliers clearly demonstrate the robustness of the modified PD classifier.

4 Multi-scale classification

For modified depth-based classification, one needs to estimate the smoothing parameter (bandwidth) involved in kernel density estimation. In Sect. 3, we used the cross-validation method for this purpose. But, using only one bandwidth pair (h_1, h_2) for classification brings in the model uncertainty. Moreover, 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 study the classification results for multiple scales of smoothing in an appropriate range and aggregate them to arrive at a new classifier, which we call the multi-scale classifier. The usefulness of multi-scale classification has been discussed in the literature both for kernel discriminant analysis and nearest neighbor classification (see, e.g., Holmes and Adams 2002, 2003; Ghosh et al. 2005, 2006). One popular way to aggregate these results (indexed by bandwidths) is to take the weighted average of the estimated posterior probabilities. Popular ensemble methods like bagging (see, e.g., Breiman 1996) and boosting (see, e.g., Schapire et al. 1998) also adopt similar ideas. 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 $\pi_1 f_1(\mathbf{x})/\pi_2 f_2(\mathbf{x})$. From this, one can obtain the estimated posterior for Class-1 as $\hat{p}_{n,h_1,h_2}(1 \mid \mathbf{x}) = e^{\xi_{n,h_1,h_2}(\mathbf{x})}/(1 + e^{\xi_{n,h_1,h_2}(\mathbf{x})})$. Aggregating these posterior estimates obtained at different values of (h_1, h_2), we arrive at the final classifier

$$d_3(\mathbf{x}) = \arg\max_{j=1,2} \; p^*_n(j|\mathbf{x}), \quad \text{where} \; p^*_n(j|\mathbf{x}) = \sum_{h_1,h_2 \in H} w_{h_1,h_2} \, \hat{p}_{n,h_1,h_2}(j|\mathbf{x}),$$

where w_{h_1,h_2} is the weight assigned to the classifier that uses h_1 and h_2 as the bandwidths of the two classes. Clearly, this aggregation depends on the bandwidth range $H = [h^l_1, h^u_1] \times [h^l_2, h^u_2]$ and the weight function w. However, if the upper and the lower bounds (h^u_j and h^l_j) of h_j (for $j = 1, 2$) satisfy (A3), then regardless of the choice of the weight function, the error rate of $d_3(\cdot)$ asymptotically converges to the Bayes risk.

Theorem 3 *Assume that h^u_j and h^l_j satisfy assumption* (A3) *for $j = 1, 2$. Furthermore, assume that f_1 and f_2 are elliptically symmetric, and they satisfy* (A1) *and* (A2). *Then, the misclassification rate of the multi-scale PD classifier $d_3(\cdot)$ converges to the Bayes risk as $min\{n_1, n_2\} \to \infty$.*

This result shows that the large sample performance of $d_3(\cdot)$ is not very sensitive to the choice of the weight function w. However, in practice, when we deal with a finite sample, one has to choose H and w appropriately. Naturally, one would use higher weights for classifiers having lower error rates, and the weight should gradually decrease as the error rate increases. Following Ghosh et al. (2006), we estimated the error rate Δ_{h_1,h_2} by the leave-one-out cross-validation method, and used the weight function $w_{h_1,h_2} = \exp\left[-\frac{1}{2}\frac{(\widehat{\Delta}_{h_1,h_2} - \widehat{\Delta}_0)^2}{\widehat{\Delta}_0(1-\widehat{\Delta}_0)/(n_1+n_2)}\right] I[\widehat{\Delta}_{h_1,h_2} \leq \min\{\pi_1, \pi_2\}]$, where $\widehat{\Delta}_0 = \min_{h_1,h_2} \widehat{\Delta}_{h_1,h_2}$. Note that $\widehat{\Delta}_0$ and $\widehat{\Delta}_0(1 - \widehat{\Delta}_0)/(n_1 + n_2)$ can be viewed as estimates for the mean and the variance of the empirical error rate of the best single scale modified PD classifier, when it is used to classify $(n_1 + n_2)$ 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 of H, we followed the method based on quantiles of the pairwise distances as described in Ghosh et al. (2006) and considered 100 equidistant values of (h_1, h_2) in that interval satisfying $h_1 = (s_1/s_2)h_2$, where s_1 and s_2 are same as in Sect. 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

To show the utility of this multi-scale (MS) approach, we analyzed the same data sets used for single scale (SS) classification. However, instead of presenting all results in another table, for better visualization, following the idea of Friedman (1994),

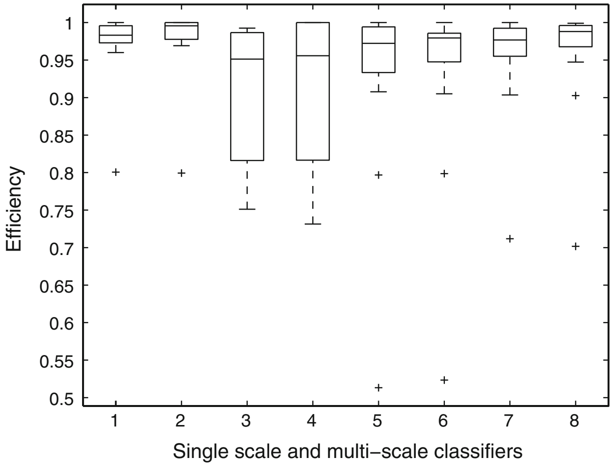

Fig. 3 Efficiencies of different single scale and multi-scale classifiers: 1 PD(SS), 2 PD(MS), 3 MD(SS), 4 MD(MS), 5 $\mathrm{MD}_{1/2}$(SS), 6 $\mathrm{MD}_{1/2}$(MS), 7 $\mathrm{MD}_{3/4}$(SS), 8 $\mathrm{MD}_{3/4}$(MS)

we used the notion of efficiency to compare the performance of different SS and MS classifiers. Since the modified HD classifier had much higher error rates compared to its corresponding versions based on PD and MD, we did not consider it for this comparison. Note that we have analyzed 15 simulated data sets (see Tables 1, 3 and 4) in this paper. For each of these data sets, we define the efficiency of the tth classifier as $\eta_t = \{\min_t e_t\}/e_t$, where e_t is the error rate of the tth classifier, and the minimization is done over all classifiers considered for comparison (i.e., SS and MS versions of MD, $\mathrm{MD}_{1/2}$, $\mathrm{MD}_{3/4}$ and PD). Note that on each data set, the best classifier has $\eta_t = 1$, and it is less than 1 for all other classifiers. A small value of η_t indicates poor performance of the tth classifier. We computed η_t for different SS and MS classifiers on all these 15 data sets, and the results are summarized using box-plots in Fig. 3, which clearly shows the importance of MS classification. For all depth-based classifiers, the overall performance of MS methods was better than of their SS counterparts. Among different depth-based classifiers, MD had higher dispersion in efficiency because of its lack of robustness. The MS version of PD had the best overall performance, followed by its SS version. The performance of $\mathrm{MD}_{1/2}$ and $\mathrm{MD}_{3/4}$ was also comparable.

5 Analysis of benchmark data sets

Now, we investigate the performance of different depth-based classifiers on four well known benchmark data sets. The vowel data was created by Peterson and Barney (1952) by a spectrographic analysis on vowels. The other three data sets are taken from the CMU data archive (http://www.statlib.cmu.edu). Since the descriptions of these data sets are available at these sources, we do not repeat them here. For the analysis of biomedical data, we ignored the observations with missing values. Though there are originally 214 observations in the glass data, 146 $(70 + 76)$ of them were from two bigger classes, and we considered those two classes only. In this data set, there are four variables with almost all values equal to zero. We ignored them and carried out our analysis with the remaining five. 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 two cases, we formed the training and the test sets by randomly partitioning the data in a such way that proportions of different classes in these two sets are as close as possible. In both these cases, we used 100 observations for training and the rest as test cases. This random partitioning was done 500 times, and the average test set error rates (over these 500 trials) of different SS and MS classifiers are reported in Table 5 along with their corresponding SE. To facilitate comparison, we also report the error rates of some standard parametric [LDA and QDA] and nonparametric [kernel discriminant analysis (KDA) and nearest neighbor classification (k-NN)] classifiers. 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 >2, it was not computationally feasible to use the grid search method. Instead, we used the Nelder–Mead algorithm available in R (see Wilcox 2005) for computing PD. The gradient descent method or the method of random selection of unit vectors (as in Maronna and Yohai 1995) can also be used for this purpose. This random search method is computationally faster, but in our experiments, it led to higher error rates than the Nelder–Mead algorithm. So, here we report the error rates of the PD classifier based on Nelder–Mead algorithm only. This algorithm retains the affine invariance property in PD calculation (see, e.g., Lagarias et al. 1998), but it makes the modified PD classifier computationally expensive compared to modified MD and modified robust MD classifiers.

In the synthetic data, since the prior probabilities of two classes are equal, we tried both maximum depth and modified depth-based classification. For maximum depth classification, error rates of PD (10.3%), MD (10.8%), $\mathrm{MD}_{1/2}$ (11.5%) and $\mathrm{MD}_{3/4}$

Table 5 Misclassification rates (in %) of different classifiers and their standard errors in real data sets

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

(11.5%) 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. In spite of that, error rates of the maximum depth classifiers were comparable to that of KDA (11.0%) and k-NN (11.7%). For all depth functions except MD, the modified depth-based classifiers performed even better, and the SS version of the modified PD classifier had the best error rate. In this vowel data, k-NN led to the best error rate (17.75%), but the error rates of all other classifiers, except LDA (25.26%) and the modified HD classifier (35.73%) was also competitive.

In the biomedical data, QDA yielded significantly lower error rate (12.57%) than KDA (16.82%) and k-NN (17.88%) methods. This gives the indication that the Gaussian model may fit the data well. Because of the validity of underlying model assumptions, modified depth-based classifiers had significantly lower error rates than KDA and k-NN. Modified MD and PD classifiers had error rates even smaller than QDA. However, LDA and QDA performed very poorly in the glass data. This indicates lack of normality in the data set. The nonparametric methods had the best error rates in this data set, but all depth-based classifiers except HD also performed well. Results on this data set show the advantage of dealing with a broader (elliptic) class.

One should also notice that the overall performance of MS classifiers was better than SS methods. It becomes more evident in the case of biomedical data and glass data, when the error rates are computed over 500 partitions. In those two data sets, almost all MS methods outperformed their SS counterparts.

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 outliers. Unlike the usual version of $L_1 D$, 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. Another major advantage of using PD is its simple relationship with Mahalanobis distance, and because of that, the PD classifier can be easily modified. The resulting modified classifier performs well for a wide variety of classification problems. While usual parametric methods like LDA and QDA work well under the normality of underlying distributions, the depth-based methods cater for a more general class of parametric models. Moreover, unlike usual nonparametric classifiers, they do not suffer from the curse of dimensionality. So, if we have a small training set in high dimension, depth-based methods are expected to outperform the nonparametric methods when the data clouds are nearly elliptic, which is quite common.

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 smaller scales of smoothing take care of the local nature of the density function and the class boundary, larger scales capture the global pattern. Incorporating these two important features in a classifier, one can expect improved performance. Using several simulated and benchmark data sets, we have amply demonstrated that in this article.

Appendix: Proofs

Proof of Theorem 1 Error rate of $d_1(\cdot)$ is given by $\Delta(d_1) = \sum_{j=1}^{J} \pi_j P\{d_1(\mathbf{X}) \neq j \mid \mathbf{X} \in j\text{th class}\}$. Under the assumptions of Theorem 1, the population version of $d_1(\cdot)$ is the Bayes classifier. So, we have

$$|\Delta(d_1) - \Delta_B| \leq \frac{1}{J} \sum_{j=1}^{J} \int \left| \prod_{i=1, i \neq j}^{J} I\left\{\frac{\mathrm{PD}(\mathbf{x}, F_{jn_j})}{\mathrm{PD}(\mathbf{x}, F_{in_i})} > 1\right\} \right. \\ \left. - \prod_{i=1, i \neq j}^{J} I\left\{\frac{\mathrm{PD}(\mathbf{x}, F_j)}{\mathrm{PD}(\mathbf{x}, F_i)} > 1\right\} \right| f_j(\mathbf{x})\, d\mathbf{x}.$$

If $f_j(\mathbf{x})$ is elliptically symmetric, it satisfies conditions $(C0)-(C3)$ of Zuo and Serfling (2000b). So, for $j = 1, \ldots, J$, we have $\sup_{\mathbf{x} \in \mathbb{R}^d} |\mathrm{PD}(\mathbf{x}, F_{jn_j}) - \mathrm{PD}(\mathbf{x}, F_j)| \overset{a.s.}{\to} 0$ as $n_j \to \infty$. Now, the convergence of $\Delta(d_1)$ to Δ_B follows from the Dominated Convergence Theorem (DCT). □

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

Proof of Lemma 1 For any $\boldsymbol{\alpha} \in \mathbb{R}^d$, since $\boldsymbol{\alpha}'\mathbf{X}$ is symmetric about $\boldsymbol{\alpha}'\boldsymbol{\mu}$, we have $\mu_F(\boldsymbol{\alpha}'\mathbf{X}) = \boldsymbol{\alpha}'\boldsymbol{\mu}$, and hence $O(\mathbf{x}, F) = \sup\limits_{\alpha : \|\alpha\| = 1} \left\{\frac{|\boldsymbol{\alpha}'\mathbf{x} - \mu_F(\boldsymbol{\alpha}'\mathbf{X})|}{\sigma_F(\boldsymbol{\alpha}'\mathbf{X})}\right\} = \sup\limits_{\alpha : \|\alpha\| = 1} \left\{\frac{|\boldsymbol{\alpha}'(\mathbf{x} - \boldsymbol{\mu})|}{\mathrm{sd}_F(\boldsymbol{\alpha}'\mathbf{X})} . \frac{\mathrm{sd}_F(\boldsymbol{\alpha}'\mathbf{X})}{\sigma_F(\boldsymbol{\alpha}'\mathbf{X})}\right\}$, where 'sd$_F$' denotes the standard deviation. Since $\mathbf{Y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ is spherically distributed, for any $\boldsymbol{\alpha} \in \mathbb{R}^d$, $\boldsymbol{\alpha}'\mathbf{Y} \overset{d}{=} ||\boldsymbol{\alpha}|| Y_1$, where Y_1 is the first component of $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ (see, e.g., Fang et al. 1989). Thus, we get $\boldsymbol{\alpha}'\mathbf{X} = \boldsymbol{\alpha}'\boldsymbol{\mu} + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{1/2}\mathbf{X} = \mu_{\boldsymbol{\alpha}} + l'_{\boldsymbol{\alpha}}\mathbf{Y} \overset{d}{=} \mu_{\boldsymbol{\alpha}} + ||l_{\boldsymbol{\alpha}}|| Y_1$, where $\mu_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}'\boldsymbol{\mu}$ and $l_{\boldsymbol{\alpha}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}$. Now, $\sigma_F(\boldsymbol{\alpha}'\mathbf{X}) = ||l_{\boldsymbol{\alpha}}|| \sigma_F(Y_1)$ and $sd_F(\boldsymbol{\alpha}'\mathbf{X}) = ||l_{\boldsymbol{\alpha}}|| sd_F(Y_1) \Rightarrow \frac{sd_F(\boldsymbol{\alpha}'\mathbf{X})}{\sigma_F(\boldsymbol{\alpha}'\mathbf{x})} = \frac{sd_F(Y_1)}{\sigma_F(Y_1)} = 1/C_F$ (say). Since C_F is free of $\boldsymbol{\alpha}$, the proof follows from the fact that $\sup_{\alpha : \|\alpha\| = 1} \{|\boldsymbol{\alpha}'(\mathbf{x} - \boldsymbol{\mu})| / sd_F(\boldsymbol{\alpha}'\mathbf{X})\} = \{(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}^{1/2}$. □

Proof of Proposition 1 Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ be the location and the scale parameters of f_j. Under elliptic symmetry of f_j, we have $f_j(\mathbf{x}) = \Gamma(d/2)(2\pi)^{-d/2} |\boldsymbol{\Sigma}_j|^{-1/2} g_j(R(\mathbf{x}, F_j)) / R(\mathbf{x}, F_j)^{d-1}$, where g_j is the p.d.f. of $R(\mathbf{x}, F_j) = \{(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\}^{1/2}$ (see, e.g., Fang et al. 1989). From Lemma 1, it follows that

$$d_B(\mathbf{x}) = \arg\max_{1\le j\le J} \pi_j f_j(\mathbf{x}) = \arg\max_{1\le j\le J} \lambda_j\theta_j\{\mathrm{O}(\mathbf{x}, F_j)\}/\{\mathrm{O}(\mathbf{x}, F_j)\}^{d-1},$$

where θ_j is the density function of $\mathrm{O}(\mathbf{x}, F_j)$ and the constant λ_j depends on F_j and π_j. Since $\mathrm{PD}(\mathbf{x}, F_j) = \{1 + \mathrm{O}(\mathbf{x}, F_j)\}^{-1}$, usual results of sampling distribution lead to a proof of Proposition 1. □

Lemma 2 *Define* $\hat{\gamma}_n(\mathbf{x}) = r^{(2)}_{n_2,h_2}(\mathbf{x})/r^{(1)}_{n_1,h_1}(\mathbf{x})$ *and recall* $\gamma(\mathbf{x}) = r^{(2)}(\mathbf{x})/r^{(1)}(\mathbf{x})$. *Under the assumptions of Theorem 2, for all* $\epsilon > 0$, *there exists* A_ϵ *such that for* $i = 1, 2$, $P(A_\epsilon \mid \mathbf{X} \in ith\ class) > 1 - \epsilon$ *and* $sup_{\mathbf{x}\in A_\epsilon}|\hat{\gamma}_n(\mathbf{x}) - \gamma(\mathbf{x})| \overset{P}{\to} 0$ *as* $min\{n_1, n_2\} \to \infty$.

Proof of Lemma 2 For $i = 1, 2$, define $\hat{\rho}_{ih_i}$ as in Sect. 2 and $\rho^*_{ih_i}(\delta) = \frac{1}{n_ih_i}\sum_{j=1}^{n_i} K\{\frac{\delta-\delta^{(i)}(\mathbf{x}_{ij})}{h_i}\}$. Also note that $\sup_{\mathbf{x}} |\hat{\rho}_{ih_i}(\hat{\delta}^{(i)}_{n_i}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))| \le \sup_{\mathbf{x}} |\hat{\rho}_{ih_i}(\hat{\delta}^{(i)}_{n_i}(\mathbf{x})) - \hat{\rho}_{ih_i}(\delta^{(i)}(\mathbf{x}))| + \sup_{\mathbf{x}} |\hat{\rho}_{ih_i}(\delta^{(i)}(\mathbf{x})) - \rho^*_{ih_i}(\delta^{(i)}(\mathbf{x}))| + \sup_{\mathbf{x}} |\rho^*_{ih_i}(\delta^{(i)}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))|$. Under elliptic symmetry of f_i, we have $\sup_{\mathbf{x}} |\hat{\delta}^{(i)}_{n_i}(\mathbf{x}) - \delta^{(i)}(\mathbf{x})| = O_P(n_i^{-1/2})$ (see, e.g., Zuo 2003). So, under (A3), it is easy to check that $\sup_{\mathbf{x}} |\hat{\rho}_{ih_i}(\hat{\delta}^{(i)}_{n_i}(\mathbf{x})) - \hat{\rho}_{ih_i}(\delta^{(i)}(\mathbf{x}))| \le M_K \sup_{\mathbf{x}} |\hat{\delta}^{(i)}_{n_i}(\mathbf{x}) - \delta^{(i)}(\mathbf{x})|/h_i^2 \overset{P}{\to} 0$ as $n_i \to \infty$, where $M_K = \sup_t |K^{'}(t)| < \infty$. Using similar arguments, we have $\sup_{\mathbf{x}} |\hat{\rho}_{ih_i}(\delta^{(i)}(\mathbf{x})) - \rho^*_{ih_i}(\delta^{(i)}(\mathbf{x}))| \overset{P}{\to} 0$ as $n_i \to \infty$. Using the properties of the kernel density estimate (see, e.g., Silverman 1986), under (A3) and uniform continuity of PD (follows from elliptic symmetry of f_i; see, e.g., Zuo 2003), we have $\sup_{\mathbf{x}} |\rho^*_{ih_i}(\delta^{(i)}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))| \overset{P}{\to} 0$ as $n_i \to \infty$. Combining the above, we get $\sup_{\mathbf{x}} |\hat{\rho}_{ih_i}(\hat{\delta}^{(i)}_{n_i}(\mathbf{x})) - \rho_i(\delta^{(i)}(\mathbf{x}))| \overset{P}{\to} 0$ as $n_i \to \infty$. Now using uniform continuity (see, e.g., Zuo 2003) and vanishing at infinity properties of PD, for any given $\epsilon > 0$, we can find $\eta = \eta(\epsilon) > 0$ such that the set $A_\epsilon = \{\mathbf{x} : \eta \le \delta^{(1)}(\mathbf{x}), \delta^{(2)}(\mathbf{x}) \le 1 - \eta\}$ has probability bigger than $1 - \epsilon$ w.r.t. probability distributions of both classes. Now, for $i = 1, 2$, it is easy to check that $\sup_{\mathbf{x}\in A_\epsilon} \left| \frac{(\hat{\delta}^{(i)}_{n_i}(\mathbf{x}))^{d-3}}{(1-\hat{\delta}^{(i)}_{n_i}(\mathbf{x}))^{d-1}} - \frac{(\delta^{(i)}(\mathbf{x}))^{d-3}}{(1-\delta^{(i)}(\mathbf{x}))^{d-1}} \right| \overset{P}{\to} 0$ and hence $\sup_{\mathbf{x}\in A_\epsilon} |r^{(i)}_{n_i,h_i}(\mathbf{x}) - r^{(i)}(\mathbf{x})| \overset{P}{\to} 0$ as $n_i \to \infty$. Also note that $\inf_{\mathbf{x}\in A_\epsilon} r^{(i)}(\mathbf{x}) > 0$ both for $i = 1, 2$, and this leads to the proof of Lemma 2. □

Lemma 3 *Define* $\Delta^{CV}_n(k) = \sum_{i=1, j\ne i}^{2} \frac{\pi_i}{n_i} \sum_{l=1}^{n_i} I\left\{\frac{r^{(j)}_{n_j,h_j}(\mathbf{x}_{il})}{r^{(i)}_{n_i,h_i}(\mathbf{x}_{il})} \ge k_i\right\}$, $\Delta(k) = \sum_{i=1, j\ne i}^{2} \pi_i\ P\left\{\frac{r^{(j)}(\mathbf{X})}{r^{(i)}(\mathbf{X})} \ge k_i \,\middle|\, \mathbf{X} \in ith\ class\right\}$, *where* $n = (n_1, n_2)$, $k_1 = 1/k$ *and* $k_2 = k$. *Also define,* $c_n = argmin_k\ \Delta^{CV}_n(k)$ *and* $c = argmin_k\ \Delta(k)$. *If* c *is unique, then under the assumptions of Theorem 2,* $c_n \overset{P}{\to} c$ *as* $min\{n_1, n_2\} \to \infty$.

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

Note that

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

Define $A_n(k_1) = \frac{1}{n_1}\sum_{i=1}^{n_1} |I\{\gamma(\mathbf{x}_{1i}) \ge k_1\} - P\{\gamma(\mathbf{X}) \ge k_1 | \mathbf{X} \in \text{ 1st class}\}|$ and $B_n(k_1) = \frac{1}{n_1}\sum_{i=1}^{n_1} \left| I\{\hat{\gamma}_n(\mathbf{x}_{1i}) \ge k_1\} - I\{\gamma(\mathbf{x}_{1i}) \ge k_1\} \right|$. Using the Glivenko–Cantelli lemma, one can show that $\sup_{k_1} |A_n(k_1)| \overset{a.s.}{\rightarrow} 0$. Under (A2), given any $\epsilon > 0$, we get a $\delta_\epsilon > 0$ such that $\sup_{k_1} |F_{\gamma,1}(k_1 + \delta_\epsilon/2) - F_{\gamma,1}(k_1 - \delta_\epsilon/2)| < \epsilon$, and A_ϵ (as in Lemma 2) such that $P(A_\epsilon \mid \mathbf{X} \in j\text{th class}) > 1 - \epsilon$ for $j = 1, 2$. Using δ_ϵ and A_ϵ, define the set $S_\epsilon = \{\mathbf{x} : |\gamma(\mathbf{x}) - k_1| > \delta_\epsilon/2\} \cap \{\mathbf{x} : \mathbf{x} \in A_\epsilon\}$. Now, we have

$$
\begin{aligned}
B_n(k_1) &= \frac{1}{n_1} \sum_{\{i:\, \mathbf{x}_{1i} \notin S_\epsilon\}} |I\{\hat{\gamma}_n(\mathbf{x}_{1i}) < k_1\} - I\{\gamma(\mathbf{x}_{1i}) < k_1\}| \\
&\quad + \frac{1}{n_1} \sum_{\{i:\, \mathbf{x}_{1i} \in S_\epsilon\}} |I\{\hat{\gamma}_n(\mathbf{x}_{1i}) < k_1\} - I\{\gamma(\mathbf{x}_{1i}) < k_1\}| \\
&\le \frac{1}{n_1} \sum_{i=1}^{n_1} I\{\mathbf{x}_{1i} \notin S_\epsilon\} + \frac{1}{n_1} \sum_{\{i:\, \mathbf{x}_{1i} \in S_\epsilon\}} |I\{\hat{\gamma}_n(\mathbf{x}_{1i}) < k_1\} - I\{\gamma(\mathbf{x}_{1i}) < k_1\}|.
\end{aligned}
$$

Note that $\frac{1}{n_1}\sum_{i=1}^{n_1} I\{\mathbf{x}_{1i} \notin S_\epsilon\} \overset{a.s.}{\rightarrow} P(\mathbf{X}_1 \notin S_\epsilon) \le P(|\gamma(\mathbf{X}_1) - k_1| \le \delta_\epsilon/2) + P(\mathbf{X}_1 \notin A_\epsilon) < 2\epsilon$ [using (A2) and Lemma 2] as $\min\{n_1, n_2\} \rightarrow \infty$. Using Lemma 2, we have $|\gamma(\mathbf{x}) - k_1| > \delta_\epsilon/2 \Rightarrow \exists\, N_0 \ge 1$ such that for all $n = (n_1, n_2)$ with $\min\{n_1, n_2\} \ge N_0$, we have $|\hat{\gamma}_n(\mathbf{x}) - k_1| > \delta_\epsilon/2$. This implies that $\frac{1}{n_1}\sum_{\{i:S(\mathbf{x}_{1i})\}} |I\{\hat{\gamma}_n(\mathbf{x}_{1i}) < k_1\} - I\{\gamma(\mathbf{x}_{1i}) < k_1\}| = 0$, and hence $B_n(k_1) \le 2\epsilon$. Now, using the same argument for $i = 2$, we get a proof of Lemma 3. □

Proof of Theorem 2 Note that $|\Delta(d_2) - \Delta_B| \leq \sum_{j=1}^{2} \int \left| \prod_{i=1, i\neq j}^{2} I\left\{ \frac{r^{(j)}_{n_j,h_j}(\mathbf{x})}{r^{(i)}_{n_i,h_i}(\mathbf{x})} \geq c_n \right\} - \prod_{i=1, i\neq j}^{2} I\left\{ \frac{r^{(j)}(\mathbf{x})}{r^{(i)}(\mathbf{x})} \geq c \right\} \right| f_j(\mathbf{x})\, d\mathbf{x}$. Using Lemmas 2, 3 and the DCT (since indicators are bounded functions), we have $|\Delta(d_2) - \Delta_B| \xrightarrow{P} 0$. Now, taking expectation w.r.t. the training sample and again using the DCT, Theorem 2 is proved. □

Proof of Theorem 3 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 the DCT. If possible, let us assume that $p^*_n(1|\mathbf{x}) \overset{P}{\nrightarrow} p(1|\mathbf{x})$. 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 $\hat{p}_{n_k,h_1^{n_k},h_2^{n_k}}(1|\mathbf{x}) \overset{P}{\nrightarrow} p(1|\mathbf{x})$. But, this sequence of bandwidths satisfy the regularity condition (A3), and hence it leads to a contradiction. □

Acknowledgements The authors would like to thank Prof. Probal Chaudhuri for useful discussions and suggestions. They are also thankful to the Associate Editor and two anonymous referees for their careful reading of the earlier versions of the manuscript, and providing them with several helpful comments and suggestions.

References

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

Cator, E. A., Lopuhaä, H. P. (2011). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. (to appear).

Chaudhuri, P., Sengupta, D. (1993). Sign tests in multidimension: Inference based on the geometry of the data cloud. *Journal of the American Statistical Association, 88*, 1363–1370.

Chen, Y., Dang, X., Peng, H., Bart, H. L. Jr (2009). Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*, 288–305.

Cover, T. M., Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*, 21–27.

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

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

Friedman, J. (1994). Flexible metric nearest neighbor classification. Technical Report LCS 113, Department of Statistics, Stanford University, California, USA. http://statistics.stanford.edu/~ckirby/techreports/LCS/LCS%20113.pdf.

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

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

Ghosh, A. K., Chaudhuri, P. (2005b). On maximum depth and related classifiers. *Scandinavian Journal of Statistics, 32*, 328–350.

Ghosh, A. K., Chaudhuri, P., Murthy, C. A. (2005). On visualization and aggregation of nearest neighbor classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1592–1602.

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

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

Hoberg, R. (2000). Cluster analysis based on data depth. In H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen, M. Schader (Eds.) *Data analysis, classification and related methods* (pp. 17–22). Berlin: Springer.

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

Holmes, C., Adams, N. (2003). Likelihood inference in nearest-neighbor classification models. *Biometrika, 90*, 99–112.

Hubert, M., Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis, 45*, 301–320.

Jornsten, R. (2004). Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis, 90*, 67–89.

Lagarias, J. C., Reeds, J. A., Wright, M. H., Wright, P. E. (1998) Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization, 9*, 112–147.

Li, J., Cuesta-Albertos, J. A., Liu, R. (2011). Nonparametric classification procedures based on DD-plot. *Submitted*. (http://personales.unican.es/cuestaj/publicaciones.html).

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

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

Maronna, R. A., Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association, 90*, 330–341.

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

Rousseeuw, P. J., Struyf, A. (1998). Computing location depth and regression depth in high dimensions. *Statistics and Computing, 8*, 193–203.

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

Schapire, R. E., Fruend, Y., Bartlett, P., Lee, W. (1998). Boosting the margin: a new explanation for the effectiveness of voting method. *Annals of Statistics, 26*, 1651–1686.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Tukey, J. (1975). Mathematics and the picturing of data. In *Proceedings of 1975 international congress of mathematicians, Vancouver* (pp. 523–531).

Tyler, D. E. (1987). A distribution free M-estimator of multivariate scatter. *Annals of Statistics, 15*, 234–251.

Vardi, Y., Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the Nationtal Academy of Sciences, USA, 97*, 1423–1426.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. London: Academic Press.

Xia, C., Lin, L., Yang, G. (2008). An extended projection data depth and its applications to discrimination. *Communications in Statistics: Theory and Methods, 37*, 2276–2290.

Zuo, Y., Serfling, R. (2000a). General notions of statistical depth function. *Annals of Statistics, 28*, 461–482.

Zuo, Y., Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *Annals of Statistics, 28*, 483–499.

Zuo, Y. (2003). Projection based depth functions and associated medians. *Annals of Statistics, 31*, 1460–1490.

Zuo, Y. (2004). Robustness of weighted L_p depth and L_p median. *Allgemeines Statistisches Archive, 88*, 1–20.