On classification based on L_p depth with an adaptive choice of p

Subhajit Dutta and Anil K. Ghosh

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, Barrackpore Trunk Road, Kolkata 700108, India. E-mail : subhajit_r@isical.ac.in, akghosh@isical.ac.in.

Abstract

In the recent past, several depth-based classifiers have been proposed in the literature for classification of multivariate data. In this article, we use L_p depth for this purpose, where p is chosen adaptively using the training data. While other depth-based classifiers have the Bayes risk consistency only under elliptic symmetry, the proposed classifier has this desirable property over a larger class of distributions. We analyze some simulated and real data sets to investigate its finite sample performance. Unlike other depth-based methods, this classifier can adopt itself to the underlying population structure. As a result, in many cases, it significantly outperforms other depth-based classifiers, especially when the underlying distributions are not elliptic.

Keywords : Bayes risk, cross-validation, data depth, kernel density estimation, l_p-symmetry, maximum likelihood estimation, misclassification rate.

1 Introduction

A depth function measures the centrality of a point $\mathbf{x}$ in $\mathbb{R}^d$ with respect to a multivariate data cloud or a multivariate probability distribution, and hence it provides a center-outward ordering of multivariate observations. Various notions of data depth (see, e.g., Liu, Parelius and Singh, 1999; Vardi and Zhang, 2000; Zuo and Serfling, 2000a) are available in the literature, and they have been used for generalizing many univariate statistical methods to the multivariate set-up. Robust estimation of multivariate location and scatter based on trimming (see, e.g., Donoho and Gasko, 1992), outlier detection (see, e.g., Chen *et. al.*, 2009), testing of multivariate statistical hypotheses (see, e.g., Liu and Singh, 1993), supervised and unsupervised classification (see, e.g., Jornsten, 2004; Hoberg and Mosler, 2006) are some examples of its wide spread applications. Christmann, Fischer and Joachims (2002) and Ghosh and Chaudhuri (2005a) used the notion of depth for linear and quadratic classification. Ghosh and Chaudhuri (2005b) introduced the notion of maximum depth classification, where an observation is assigned to the class with respect to which it has the maximum data depth. They also developed a generalized classifier based on half-space depth (HD) (see, e.g., Tukey, 1975), which performs better than the maximum depth classifier in a wide variety of classification problems. Later, Dutta and Ghosh (2011) used a robust version of Mahalanobis depth (MD) and projection depth (PD) (see, e.g., Zuo and Serfling, 2000a) for this purpose, and demonstrated their superiority over the classifier based on HD. They also proved the Bayes risk consistency of their proposed classifiers (i.e., the convergence of their misclassification rates to the

Bayes risk) when the underlying distributions are elliptically symmetric (see, e.g., Fang, Kotz and Ng, 1989). López-Pintado and Romo (2006) and Cuevas and Fraiman (2009) introduced depth-based methods for classification of functional data.

Note that in the case of an elliptically symmetric distribution, for any affine invariant depth function, the density function turns out to be a function of depth (see Zuo and Serfling, 2000b). So, if the competing population distributions are all elliptically symmetric, the Bayes classifier can also be expressed as a function of depths, and hence it is natural to develop depth-based classification tools. However, for most of the existing notions of depth, this functional relationship between depth and density functions is not known explicitly, and that hinders the construction of meaningful classification rules based on depth. Of course, this relationship is straight forward for MD (see Hartikainen and Oja, 2006), and it has also been derived for HD (see Ghosh and Chaudhuri, 2005b) and PD (see Dutta and Ghosh, 2011). So, these three notions of data depth have been used to develop generalized depth-based classifiers (see, e.g., Dutta and Ghosh, 2011), which have the Bayes risk consistency when the underlying population densities are elliptically symmetric (i.e., l_2-symmetric after standardization).

Now, one may be curious to know how these depth-based classifiers perform when the underlying densities are not necessarily elliptic. For instance, let us consider the case when they are l_p-symmetric (after standardization) for some positive $p(\neq 2)$. MD of an observation $\mathbf{x}$ with respect to a distribution F is defined as $\text{MD}(\mathbf{x}, F) = \{1 + O_{MD}(\mathbf{x}, F)\}^{-1}$, where $O_{MD}(\mathbf{x}, F) = (\mathbf{x} - \boldsymbol{\mu}_F)'\boldsymbol{\Sigma}_F^{-1}(\mathbf{x} - \boldsymbol{\mu}_F)$ for $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ being the location and the scale parameters of F. From this definition, it is clear that irrespective of the underlying density function, MD contours will always be spherical, and hence, the density cannot be a function of MD. Dutta, Ghosh and Chaudhuri (2011) proved that for any l_p-symmetric distribution with $p \neq 2$, the density cannot be a function of HD as well. PD of an observation $\mathbf{x}$ with respect to F is defined as $\text{PD}(\mathbf{x}, F) = \{1 + O_{PD}(\mathbf{x}, F)\}^{-1}$, where $O_{PD}(\mathbf{x}, F) = \sup_{\boldsymbol{\alpha}}\{|\boldsymbol{\alpha}'\mathbf{x} - \text{median}_F(\boldsymbol{\alpha}'\mathbf{X})|/\text{MAD}_F(\boldsymbol{\alpha}'\mathbf{X})\}$ and MAD is median absolute deviation about median. Note that if we replace the median by the mean and the MAD by the standard deviation, we get back MD. So, being a robust version of MD, PD is likely to have properties somewhat similar to MD. Since PD contours are convex (see Lemma 1 in Appendix), density contours cannot match PD contours when $p < 1$. For the case $p = 1$ and $p = \infty$, one can choose three points $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_\theta = \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2$ $(0 < \theta < 1)$ on the same l_p contour such that $\text{PD}(\mathbf{x}_\theta, F) > \text{PD}(\mathbf{x}_1, F) = \text{PD}(\mathbf{x}_2, F)$. In particular, in $\mathbb{R}^2$, we can choose $\mathbf{x}_1 = (1, 0)$, $\mathbf{x}_2 = (0, 1)$ and any $\theta \in (0, 1)$ for $p = 1$. For $p = \infty$, instead of $(0, 1)$, $\mathbf{x}_2 = (1, 1)$ can be chosen. Figure 1 shows HD and PD contours (indicated using `bold curves`) computed based on 2000 observations from two l_p-symmetric densities (density contours are shown using `dotted curves`) with $p = 1$ and $p = 5$. This figure clearly suggests that the density cannot be a function of depth in either of these cases. Therefore, in such cases, the Bayes rule will not be a function of these depths, and the classifiers based on PD, MD and HD may fail to yield satisfactory results. But, we can overcome this limitation by using L_p depth (see Zuo and Serfling, 2000a) with an

appropriately chosen value of p. In that case, depth contours will coincide with density contours, and the classifier based on L_p depth is expected to outperform other depth-based classifiers.

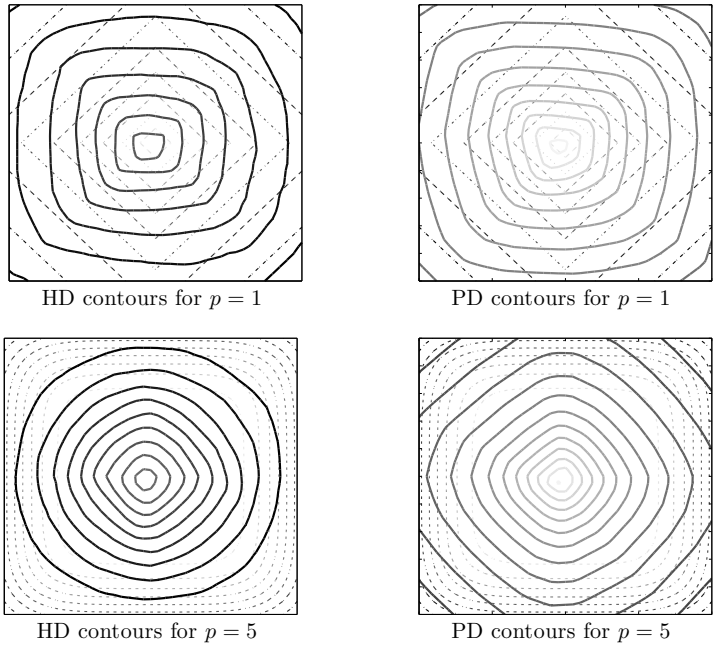

Figure 1. Density contours (`dotted curves`) and its corresponding HD and PD contours (`bold curves`), for $p = 1$ and $p = 5$.

2 Classification using L_p depth

Before using L_p depth for classification, one has to estimate the value of p for each of the J competing populations. Here, we assume that each of the population distributions is l_p-symmetric (after standardization) for some $p > 0$ (which can be different for different populations), and estimate the value of p which fits the data well. We start by discussing our methodology for a single population, and then extend it to the classification framework. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are n independent observations from a common distribution with density f. If $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the location and the scale parameters of the distribution, respectively, and the density f is of the form $f(\mathbf{x}, p) = \psi(\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_p)$ for some $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ and $p > 0$, then f can be expressed as

$$f(\mathbf{x}, p) = |\boldsymbol{\Sigma}|^{-1/2} \, \frac{p^{d-1}\Gamma(d/p)}{2^d\{\Gamma(1/p)\}^d} \frac{g_p(r_p(\mathbf{x}))}{r_p(\mathbf{x})^{d-1}}, \tag{1}$$

where $r_p(\mathbf{x}) = \|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|_p$, g_p is the density of $r_p(\mathbf{X})$ when $\mathbf{X} \sim f$ (see Lemma 2 in Appendix), and for any $\mathbf{z} = (z_1, \cdots, z_d) \in \mathbb{R}^d$ and $p > 0$, $\|\mathbf{z}\|_p$ is defined as $(|z_1|^p + \cdots + |z_d|^p)^{1/p}$. We estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the data, and those estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are used to define $\hat{r}_p(\mathbf{x}_i) = \|\hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})\|_p$ for $i = 1, 2, \ldots, n$. Usual moment-based estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be used for this purpose. But, here we consider MCD estimates (see, e.g., Rousseeuw and Leroy, 1987) to make it robust, and use

$\alpha = 0.75$ as suggested in Hubert and van Driessen (2004) for better finite sample performance. Since $\hat{\mathbf{\Sigma}}^{-1/2}$ is involved in the computation of the $\hat{r}_p(\mathbf{x}_i)$s, MCD estimates will not make it affine invariant. In order to make the $\hat{r}_p(\mathbf{x}_i)$s affine invariant, one can use the transformation re-transformation technique discussed in Chakraborty and Chaudhuri (1996). Using $\hat{r}_p(\mathbf{x}_1), \hat{r}_p(\mathbf{x}_2), \ldots, \hat{r}_p(\mathbf{x}_n)$ as sample observations, we estimate g_p using the method of kernel density estimation (see, e.g., Silverman, 1998). This univariate density estimate is given by $\hat{g}_{p,h}(r) = \frac{1}{nh}\sum_{i=1}^{n} K\left[\frac{r-\hat{r}_p(\mathbf{x}_i)}{h}\right]$, where K is the kernel function, and h is the associated bandwidth parameter. Throughout this article, we use the Gaussian kernel $K(t) = (2\pi)^{-1/2}e^{-t^2/2}$ for our analysis, and h is chosen using the bandwidth selection method proposed in Sheather and Jones (1991). So, the estimate of $f(\mathbf{x}, p)$ is given by

$$\hat{f}_h(\mathbf{x}, p) = |\hat{\mathbf{\Sigma}}|^{-1/2}\ \frac{p^{d-1}\Gamma(d/p)}{2^d\{\Gamma(1/p)\}^d}\frac{\hat{g}_{p,h}(\hat{r}_p(\mathbf{x}))}{[\hat{r}_p(\mathbf{x})]^{d-1}}. \tag{2}$$

To estimate p, one can consider the estimated joint likelihood function $\mathcal{L}_p(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \prod_{i=1}^{n} \hat{f}_h(\mathbf{x}_i, p)$, and choose the value of p by maximizing $\mathcal{L}_p(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ or $\log \mathcal{L}_p(\mathbf{x}_1, \cdots, \mathbf{x}_n)$. However, if $\hat{r}_p(\mathbf{x})$ is close to zero or infinity, $|\log \hat{f}_h(\mathbf{x}, p)|$ will be very influential. So, we consider only those $\mathbf{x}_i$s for which $\hat{r}_p(\mathbf{x}_i)$ lie that between ζ_{1n} and ζ_{2n} $(0 < \zeta_{1n} < \zeta_{2n} < \infty)$, and find $\hat{p}$ (the estimated value of p) by maximizing $l_p(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \sum_{\{i\ :\ r_p(\mathbf{x}_i)\in[\zeta_{1n},\zeta_{2n}]\}} \log \hat{f}_h(\mathbf{x}_i, p)$. The following theorem suggests a suitable choice for $[\zeta_{1n}, \zeta_{2n}]$ and shows the consistency of $\hat{p}$ under appropriate regularity conditions.

Theorem 1 : *Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are i.i.d. observations from $f(\mathbf{x}, p_0)$, which is of the form $f(\mathbf{x}, p_0) = \psi(\|\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|_{p_0})$ for some $\boldsymbol{\mu}$, $\mathbf{\Sigma}$, $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ and $p_0 \geq 1$. For any $p \geq 1$, define $\zeta_{1n} = \zeta_1(p)$ and $\zeta_{2n} = \zeta_2(p)$ as the α_{1n}-th and α_{2n}-th quantile of g_p, where $\alpha_{1n} \to 0$ and $\alpha_{2n} \to 1$ as $n \to \infty$. Also, assume that the following holds*

$$\textit{(C1) } \sqrt{n}\ \zeta_{1n} \to \infty \textit{ as } n \to \infty \textit{ and (C2) } n^{1/10} \inf_{\{r_p(\mathbf{x})\in\ [\zeta_{1n},\zeta_{2n}]\}} g_p(r_p(\mathbf{x})) \to \infty \textit{ as } n \to \infty.$$

Then, $\hat{p}_n = \arg\max_{p\geq 1} l_p(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ converges in probability to p_0 as $n \to \infty$.

Note that for any distribution function with density g_p, one can always choose ζ_{1n} and ζ_{2n}, or equivalently, α_{1n} and α_{2n} such that $(C1)$ holds. Moreover, if g_p is bounded away from zero and infinity, $(C2)$ holds for any choice of α_{1n} and α_{2n}. So, these conditions usually hold in practice. Theorem 1 tells us about the consistency of $\hat{p}$ when $p_0 \geq 1$ and $l_p(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ is maximized over the range $[1, \infty)$. Densities with $p < 1$ are rare in practice, and later we will carry out simulation studies to investigate the performance of our method in those cases.

2.1 Maximum L_p depth classifier

The maximum L_p depth classifier classifies an observation to the class with respect to which it has the maximum L_p depth ($\mathrm{L}_p\mathrm{D}$). L_p depth of an observation $\mathbf{x}$ with respect to a distribution F is defined as $\mathrm{L}_p\mathrm{D}(\mathbf{x}, F) = [1 + r_p(\mathbf{x})]^{-1}$, where $r_p(\mathbf{x}) = \|\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_p$ for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ being the location and the scale parameters of F. To construct the empirical version of $\mathrm{L}_p\mathrm{D}$, we estimate p and $r_p(\mathbf{x})$ from the data following the method described in the preceding subsection. For computation

of $l_p(\mathbf{x}_1, \cdots, \mathbf{x}_n)$, in our data analysis, we consider all observations with $10^{-30} \leq r_p(\mathbf{x}) \leq 10^{30}$. Suppose that there are J competing classes with distributions $F_1, F_2, \ldots, F_J$, and for the j-th ($j = 1, 2, \ldots, J$) class, these estimates are denoted by $\hat{p}_j$ and $\hat{r}_{\hat{p}_j}(\mathbf{x})$, respectively. Then, the empirical version of $\mathrm{L}_p\mathrm{D}(\mathbf{x}, F_j)$ is given by $[1 + \hat{r}_{\hat{p}_j,j}(\mathbf{x})]^{-1}$. So, the maximum L_p depth classifier classifies an observation $\mathbf{x}$ to the j-th class if $\hat{r}_{\hat{p}_j,j}(\mathbf{x}) < \hat{r}_{\hat{p}_i,i}(\mathbf{x})$ for all $i \neq j$. However, we know that maximum depth classification is particularly useful when the population distributions differ only in their location (see, e.g., Ghosh and Chaudhuri, 2005b). So, in that case, it is more reasonable to use a common p for all classes. If $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are observations from the j-th ($j = 1, \ldots, J$) class, this common value of p can be estimated by maximizing $\sum_{j=1}^{J} l_p(\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j})$. The resulting maximum depth classifier is given by $\mathsf{d}_1(\mathbf{x}) = \arg\min_{1 \leq j \leq J} \hat{r}_{\hat{p},j}(\mathbf{x})$, where $\hat{p}$ denotes the estimated value of p.

Theorem 2 : *Suppose that the population density functions $f_1, \ldots, f_J$ are unimodal, and f_j ($1 \leq j \leq J$) is of the form $f_j(\mathbf{x}, p_0) = \psi(\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_j)\|_{p_0})$ for some $p_0 \geq 1$. If the prior probabilities of the competing classes are equal, under conditions $(C1)$ and $(C2)$, the misclassification rate of $\mathsf{d}_1(\mathbf{x})$ converges to the Bayes risk as $min\{n_1, \ldots, n_J\} \to \infty$.*

To evaluate the performance of this maximum $\mathrm{L}_p\mathrm{D}$ classifier, we used some simulated datasets consisting of bivariate observations from two l_p-symmetric distributions. These two distributions had different location parameters $\boldsymbol{\mu}_1 = (0, 0)$ and $\boldsymbol{\mu}_2$ [we considered two different choices for $\boldsymbol{\mu}_2$, $(1, 1)$ and $(1, 2)$], but the same scatter matrix $\boldsymbol{\Sigma} = \mathbf{I_2}$ (the 2×2 identity matrix), same value of p and the same functional form $\psi(x) = C_p \exp(-\|\mathbf{x}\|_p^p)$, where C_p is the normalizing constant. We used four different values of p ($p = 1/2, 1, 2, 8$), and in each case, taking equal number of observations from the two classes, we generated training and test sets of sizes 400 and 1000, respectively. This experiment was repeated 100 times, and the average error rate of the maximum $\mathrm{L}_p\mathrm{D}$ classifier was computed over these 100 trials. To compare its performance with other maximum depth classifiers, following Ghosh and Hall (2008), we computed regret functions (difference between the average error rate and the Bayes risk) for different classifiers. Table 1 presents the corresponding regret ratios for maximum depth classifiers based on HD, PD and robust MD computed using MCD estimates with $\alpha = 0.75$. Regret ratio (η_t) of the classifier t is given by ratio of the regret of that classifier and that of the maximum $\mathrm{L}_p\mathrm{D}$ classifier. Clearly, $\eta_t < 1 (> 1)$ implies that the classifier is better (worse) than the maximum $\mathrm{L}_p\mathrm{D}$ classifier, and the deviation from 1 gives an idea of how better (worse) it is. Throughout this section, we used equal prior probabilities for the two classes.

Table 1 : Regret ratios of different maximum depth classifiers

$\boldsymbol{\mu}_2$	(1,1)				(1,2)			
p	1/2	1	2	8	1/2	1	2	8
HD	2.28	53.01	30.19	25.94	1.86	8.49	56.96	8.18
PD	1.81	1.45	2.97	6.31	1.33	1.90	5.22	19.98
MD	1.46	1.26	0.96	1.17	1.37	1.81	0.94	9.15

Table 1 shows that the overall performance of the $\mathrm{L}_p\mathrm{D}$ classifier was better than other maximum

depth classifiers. Though the performance of robust MD classifier was comparable with the L_pD classifier, classifiers based on PD and HD had much higher regret ratios even in the case $p = 2$, where all these depth-based classifiers have the Bayes risk consistency. It is interesting to note that in the case of $\boldsymbol{\mu}_2 = (1, 1)$, not only for $p = 2$, but also in all other cases, the robust MD classifier had regret ratios close to 1. Notice that in this case, the conditions $\|\mathbf{x} - \boldsymbol{\mu}_1\|_2 < \|\mathbf{x} - \boldsymbol{\mu}_2\|_2$ and $\|\mathbf{x} - \boldsymbol{\mu}_1\|_p < \|\mathbf{x} - \boldsymbol{\mu}_2\|_p$ are equivalent for any $p > 0$. So, the maximum depth classifiers based on the robust version of MD and L_pD are expected to have similar performance. But, this was not the case for $\boldsymbol{\mu}_2 = (1, 2)$, and there the robust MD classifier had considerably poor performance than the L_pD classifier in some situations.

2.2 Generalized L_p depth classifier

Maximum depth classifiers are particularly useful when the population distributions have the same prior, and they differ only in their location. But, in practice, the population distributions may have different prior probabilities, and they can also have different scatters as well as shapes. In such cases, the maximum depth classifier may lead to poor performance, and it can even perform as worse as random guessing. Therefore, in such situations, one needs to develop a new depth-based classification rule. Here we develop the generalized L_pD classifier with this main objective.

Suppose that we have observations from J competing populations, and $\hat{p}_j$ is the estimated value of p_j $(1 \leq j \leq J)$. Then, using (2), one can estimate $f_j(\mathbf{x}, p_j)$ by $\hat{f}_{h_j}(\mathbf{x}, \hat{p}_j)$. Note that irrespective of the dimension of the measurement vector, here we need one-dimensional kernel density estimation, and this helps to get rid of the curse of dimensionality that one faces in high dimensional nonparametric density estimation. Such depth-based density estimation and its advantages were also discussed in Fraiman, Liu and Mechole (1997). In a two-class problem, we consider a classifier of the form

$$\mathsf{d}_2(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{f}_{h_1}(\mathbf{x}, \hat{p}_1)/\hat{f}_{h_2}(\mathbf{x}, \hat{p}_2) > k \\ 2 & \text{otherwise,} \end{cases} \tag{3}$$

and estimate k by minimizing the leave-one-out cross-validation error rate (see, e.g., Hastie, Tibshirani and Friedman, 2009). Note that under appropriate regularity conditions, this estimate $\hat{f}_{h_j}(\mathbf{x}, \hat{p}_j)$ converges to a scalar multiple of $f_j(\mathbf{x}, p_j)$ (see Lemma 4 in Appendix). So, instead of choosing $k = \pi_2/\pi_1$, it is better to choose k using the cross-validation technique. If there are more than two classes, we can adopt the pairwise classification approach, and the results of all pairwise classifications can be combined using the method of majority voting. Recall that computation of $\hat{f}_{h_j}(\mathbf{x}, \hat{p}_j)$ requires the bandwidth h_j to be selected, and here we have used the Sheather-Jones bandwidth for this purpose.

Theorem 3 : *Suppose that for all $j = 1, 2, \ldots, J$, the density function f_j is of the form $f(\mathbf{x}, p_j) = \psi_j(\|\boldsymbol{\Sigma}_j^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_j)\|_{p_j})$ for some $p_j \geq 1$. Under conditions $(C1)$ and $(C2)$, the error rate of the generalized L_p depth classifier $\mathsf{d}_2(\mathbf{x})$ converges to the Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.*

To compare the performance of the proposed method with other generalized classifiers based on robust MD, HD and PD, we carried out a simulation study. To keep our examples simple, here we restricted ourselves to two-class problems in two dimensions. In all these cases, the density function of the j-th ($j = 1, 2$) class was considered to be of the form $f_j(\mathbf{x}) = C_{p_j}|\mathbf{\Sigma}_j|^{-1/2} \exp\{-\|\mathbf{\Sigma}_j^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_j)\|_{p_j}^{p_j}\}$, where C_{p_j} is a normalizing constant. Note that for varying choices of $(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1, p_1)$ and $(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2, p_2)$, one gets different types of classification problems (see Table 2). In each of these cases, taking equal number of observations from the two competing classes, we generated training and test sets of size 400 and 1000, respectively. Each experiment was repeated 100 times as before. As in the case of maximum depth classification, here also we calculated regret ratios of other depth-based classifiers with respect to the generalized $\mathrm{L}_p\mathrm{D}$ classifier, and they are reported in Table 2. Throughout this section prior probabilities of the two classes are taken to be equal.

First, we consider the cases where the two populations differ only in their scales (i.e., $\mathbf{\Sigma}_1 = \mathbf{I}_2$ and $\mathbf{\Sigma}_2 = 9\mathbf{I}_2$). We carried out this experiment for four different values of p. As expected, for $p = 2$, MD had the best performance, but in other three cases, $\mathrm{L}_p\mathrm{D}$ turned out to be the winner. For $p = 1/2$ and $p = 1$, the performance of MD and PD was quite comparable to that of the generalized $\mathrm{L}_p\mathrm{D}$ classifier, but in the case $p = 8$, $\mathrm{L}_p\mathrm{D}$ clearly outperformed its competitors.

Table 2 : Regret ratios of generalized depth-based classifiers.

Difference is scales					Difference in shapes				
(p_1, μ_1, Σ_1)	(p_2, μ_2, Σ_2)	HD	PD	MD	(p_1, μ_1, Σ_1)	(p_2, μ_2, Σ_2)	HD	PD	MD
$(1/2, (0,0), I_2)$	$(1/2, (0,0), 9I_2)$	2.11	1.08	1.22	$(1, (0,0), I_2)$	$(2, (0,0), I_2)$	2.91	1.51	1.32
$(1, (0,0), I_2)$	$(1, (0,0), 9I_2)$	2.59	1.13	1.01	$(2, (0,0), I_2)$	$(4, (0,0), I_2)$	2.23	2.12	1.30
$(2, (0,0), I_2)$	$(2, (0,0), 9I_2)$	5.88	1.50	0.79	$(1, (0,0), I_2)$	$(8, (0,0), I_2)$	3.40	3.04	1.93
$(8, (0,0), I_2)$	$(8, (0,0), 9I_2)$	5.84	3.93	2.06	$(1/2, (0,0), I_2)$	$(16, (0,0), I_2)$	4.88	1.84	1.43
Non-truncated vs. Truncated					Difference in shapes and locations				
(p_1, μ_1, Σ_1)	(p_2, μ_2, Σ_2)	HD	PD	MD	(p_1, μ_1, Σ_1)	(p_2, μ_2, Σ_2)	HD	PD	MD
$(1/2, (0,0), I_2)$	$T(1/2, (0,0), I_2)$	3.47	2.37	1.94	$(1, (0,0), I_2)$	$(2, (1,1), I_2)$	2.52	2.49	1.75
$(1, (0,0), I_2)$	$T(1, (0,0), I_2)$	3.82	2.05	1.49	$(2, (0,0), I_2)$	$(4, (1,1), I_2)$	1.52	1.40	1.14
$(2, (0,0), I_2)$	$T(2, (0,0), I_2)$	3.71	1.66	0.98	$(1, (0,0), I_2)$	$(8, (1,1), I_2)$	14.62	2.85	1.67
$(8, (0,0), I_2)$	$T(8, (0,0), I_2)$	3.02	2.34	1.52	$(1/2, (0,0), I_2)$	$(16, (1,1), I_2)$	15.35	2.08	1.42

Next we consider some examples where the first population is kept unchanged, but a truncated version of it is considered as the second population. We chose four different values of p, and in each case, the second population had observations $\mathbf{x}$ with $\|\mathbf{x}\|_p \geq c_p$, where c_p is chosen such that the Bayes risk is 0.25. For $p = 2$, MD and $\mathrm{L}_p\mathrm{D}$ had comparable performance, but in all other cases, the performance of the generalized $\mathrm{L}_p\mathrm{D}$ classifier was much better than its competitors.

We also consider the cases, where the two populations differ only in their shapes (i.e., $p_1 \neq p_2$). We consider four different choices for (p_1, p_2), and in all these cases, the generalized $\mathrm{L}_p\mathrm{D}$ classifier yielded the best performance. We also observed the same phenomenon when the two populations differed in their locations as well as shapes (see Table 2). These examples clearly show the superiority of the generalized $\mathrm{L}_p\mathrm{D}$ classifier over other classification tools based on data depth.

The HD classifier had the highest regret ratios in all examples.

3 Results from the analysis of benchmark datasets

We analyze six benchmark data sets for further illustration of the proposed classifier. All these data sets except the hemophilia data are taken either from the UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/`) or from the Statlib Datasets Archive at CMU (`http://lib.stat.cmu.edu/datasets/`). The hemophilia data set is taken from Johnson and Wichern (1992). Since the description of these data sets are available at these sources, we do not repeat them here. For the synthetic data, the training and the test sets are well specified. In all other cases, we formed these sets by randomly partitioning the data in such a way that the proportion of different classes in these two sets are as close as possible. In cases of glass, biomedical and diabetes data sets, 100 observations were used to constitute the training sample, whereas in the case of hemophilia data, we used training samples of size 50. The blood transfusion data consist of a reasonably large number of observations, and in that case, we split the data set into two equal halves. In all these cases, 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 generalized depth-based classifiers are reported in Table 3 along with their corresponding standard errors. For MD and L_pD, we used both the robust (MCD with $\alpha = .75$) and the non-robust (moment based) estimates of the location and the scatter parameters, and the error rates for both these methods are reported in Table 3. Note that under some further moment conditions on the underlying distribution, the Bayes risk consistency of the generalized L_pD classifier holds even when the moment based estimates are plugged in. To facilitate comparison, error rates are also reported for two parametric (linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)) and two nonparametric (kernel discriminant analysis (KDA) and nearest neighbor classifier (k-NN)) classifiers (see, e.g., Hastie *et. al.* (2009) for a discussion on these classifiers). Since the classifier based on HD had the highest error rates in almost all cases in our simulation studies, we do not report them here. In the glass data set, though there were nine variables, four of them had almost all values equal to zero. Considering those variables to be of less importance for classification, we carried out our analysis based on the remaining five. Though there are originally 214 observations in this data set, 146 of them were from two bigger classes, and here we have considered those two classes only. In the case of biomedical data set, we ignored the 15 observations with missing values, and carried out our analysis using the remaining 194 observations. In the case of blood transfusion data, following Li, Cuesta-Albertos and Liu (2011), we removed one of the two linearly dependent variables from the data set. Throughout this section, sample proportions of different classes have been used as their prior probabilities.

The overall performance of the generalized L_pD classifier was fairly satisfactory. Except for the biomedical data, in all other cases, robust versions of L_pD and MD classifiers based on MCD

Table 3 : Average misclassification rates (in %) of different classifiers and their standard errors.

Methods → Datasets ↓	LDA	QDA	k-NN	KDA	PD	MD		L_pD	
						Moment	MCD	Moment	MCD
Synthetic	10.80	10.20	11.70	11.00	10.70	10.20	10.60	9.60	10.70
Hemophilia	15.22	15.47	15.79	15.58	18.65	15.84	17.13	15.39	16.43
	(0.27)	(0.26)	(0.30)	(0.28)	(0.35)	(0.30)	(0.32)	(0.32)	(0.32)
Glass	30.69	36.13	22.78	21.94	24.95	26.80	24.80	27.64	24.75
	(0.25)	(0.25)	(0.24)	(0.23)	(0.27)	(0.26)	(0.29)	(0.29)	(0.26)
Biomedical	15.64	12.57	17.81	16.79	12.41	12.35	14.48	12.68	15.11
	(0.12)	(0.12)	(0.14)	(0.15)	(0.13)	(0.14)	(0.15)	(0.15)	(0.15)
Diabetes	10.46	9.39	10.04	11.09	14.01	8.22	11.49	9.39	11.92
	(0.18)	(0.18)	(0.18)	(0.19)	(0.22)	(0.18)	(0.22)	(0.21)	(0.27)
Blood	23.08	22.61	27.69	24.38	23.99	22.75	22.17	22.30	22.06
Transfusion	(0.03)	(0.05)	(0.09)	(0.03)	(0.13)	(0.07)	(0.08)	(0.07)	(0.07)

estimates had lower error rates than the PD classifier, and in some cases (hemophilia, diabetes and blood transfusion datasets), this difference was statistically significant at 5% level when the usual t-test was used. The performance of the non-robust version of L_pD classifier based on moment based estimates is reported here to compare it with non-robust parametric methods like LDA and QDA. This version of L_pD classifier performed better than LDA in all cases except for the hemophilia data, where they had comparable error rates. It also had significantly better performance than QDA in glass data and blood transfusion data, while in all other cases, there was no significant difference between their error rates. In the biomedical data as well as in the blood transfusion data, both versions of the L_pD classifier had significantly lower error rates than the nonparametric methods. The moment based version yielded lower error rates also in diabetes, hemophilia and synthetic data sets. Only in the case of glass data, nonparametric methods outperformed LDA, QDA and all depth-based classifiers.

Acknowledgment

The authors would like to thank Professor Probal Chaudhuri for useful discussions and suggestions.

Appendix

Lemma 1 : Projection depth (PD) contours are convex.

Proof of Lemma 1 : Define $S_F(\delta) = \{\mathbf{x} : O_{PD}(\mathbf{x}, F) \leq \delta\}$, and choose two points $\mathbf{x}_0$ and $\mathbf{x}_1$ in $S_F(\delta)$ such that $O_{PD}(\mathbf{x}_0, F) = O_{PD}(\mathbf{x}_1, F) = \delta$. If possible, assume $S_F(\delta)$ is not convex, and $\mathbf{x}_\lambda = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_0 \notin S_F(\delta)$ for some $\lambda \in (0,1)$. But, using the definition of $O_{PD}(\mathbf{x}, F)$, one can check that $O_{PD}(\mathbf{x}_\lambda, F) \leq \lambda O_{PD}(\mathbf{x}_1, F) + (1-\lambda) O_{PD}(\mathbf{x}_0, F) = \delta$. So, this leads to a contradiction.

□

Lemma 2 : If $f(\mathbf{x}, p)$ is of the form $f(\mathbf{x}, p) = \psi(\|\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_p)$ for some $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ and $p > 0$, it can also be expressed as

$$f(\mathbf{x}, p) = |\mathbf{\Sigma}|^{-1/2} \frac{p^{d-1}\Gamma(d/p)}{2^d\{\Gamma(1/p)\}^d} \frac{g_p(r_p(\mathbf{x}))}{r_p(\mathbf{x})^{d-1}},\ r_p(\mathbf{x}) > 0$$

where $r_p(\mathbf{x}) = \|\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_p$, g_p is the density of $r_p(\mathbf{X})$ and $\mathbf{X} \sim f(\cdot, p)$.

Proof of Lemma 2 : Define $\mathbf{Y} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. If f_0 denotes the density of $\mathbf{Y}$, it is easy to see that $f_0(\mathbf{y}) = |\mathbf{\Sigma}|^{1/2} f(\mathbf{y}, p) = |\mathbf{\Sigma}|^{1/2}\psi(\|\mathbf{y}\|_p)$. Now, following the proof of Lemma 1.4 in Fang, Kotz and Ng (1989), one can show that for any non-negative measurable function ξ, we have

$$\int_{\mathbb{R}^d} \xi(\|\mathbf{y}\|_p)\ d\mathbf{y} = \frac{2^d\{\Gamma(1/p)\}^d}{p^d\Gamma(d/p)} \int_0^\infty \xi(u^{1/p}) u^{d/p-1} du.$$

So, for any non-negative measurable function ϕ, defining $\xi = \phi.\psi$ we get

$$\begin{aligned} E[\phi(\|\mathbf{Y}\|_p)] &= |\mathbf{\Sigma}|^{1/2} \int \phi(\|\mathbf{y}\|_p)\psi(\|\mathbf{y}\|_p)\ d\mathbf{y} \\ &= |\mathbf{\Sigma}|^{1/2} \frac{2^d\{\Gamma(1/p)\}^d}{p^d\Gamma(d/p)} \int_0^\infty \phi(u^{1/p})\ \psi(u^{1/p})\ u^{d/p-1}\ du \\ &= |\mathbf{\Sigma}|^{1/2} \frac{2^d\{\Gamma(1/p)\}^d}{p^{d-1}\Gamma(d/p)} \int_0^\infty \phi(r)\ \psi(r)\ r^{d-1}\ dr \text{ (letting } r = u^{1/p}). \end{aligned}$$

So, g_p, the density of $r_p(\mathbf{X}) = \|\mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})\|_p$ turns out to be of the form

$$g_p(r) = |\mathbf{\Sigma}|^{1/2} \frac{2^d\{\Gamma(1/p)\}^d}{p^{d-1}\Gamma(d/p)}\ \psi(r)\ r^{d-1} \Rightarrow \psi(r_p(\mathbf{x})) = |\mathbf{\Sigma}|^{-1/2} \frac{p^{d-1}\Gamma(d/p)}{2^d\{\Gamma(1/p)\}^d}\ \frac{g_p(r_p(\mathbf{x}))}{r_p(\mathbf{x})^{d-1}}.\ \square$$

Lemma 3 : If f is l_p-symmetric, then for any $p \geq 1$, $\sup_{\mathbf{X}\in\mathbb{R}^d} |\hat{r}_p(\mathbf{x}) - a_{p_0}^{-1/2} r_p(\mathbf{x})| \xrightarrow{P} 0$ as $n \to \infty$ for some $a_{p_0} > 0$. If $g_p^{(a)} (= a_{p_0}^{1/2} g_p)$ denotes the density of $a_{p_0}^{-1/2} r_p(\mathbf{X})$, then we have

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \xrightarrow{P} 0 \text{ as } n \to \infty.$$

Proof of Lemma 3 : Note that under l_p-symmetry of the distribution, we have $\hat{\boldsymbol{\mu}} \xrightarrow{P} \boldsymbol{\mu}$ and $\hat{\mathbf{\Sigma}} \xrightarrow{P} a_{p_0}\mathbf{\Sigma}$ (see Cator and Lopuhaä, 2011 for the convergence property of MCD estimates), where $a_{p_0} > 0$ and it depends on the underlying density function $f(\mathbf{x}, p_0)$. So, for $p \geq 1$, we have

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{r}_p(\mathbf{x}) - a_{p_0}^{-1/2} r_p(\mathbf{x})| \xrightarrow{P} 0 \text{ as } n \to \infty. \tag{4}$$

Now, using the triangle inequality, we have

$$\begin{aligned} &|\hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \\ &\leq |\hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - \hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| + |\hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))|. \end{aligned} \tag{5}$$

We consider the first term in (5), and again using the triangle inequality, we get

$$|\hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - \hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \leq \frac{1}{nh} \sum_{i=1}^{n} \left| K\left[\frac{\hat{r}_p(\mathbf{x}) - \hat{r}_p(\mathbf{x}_i)}{h}\right] - K\left[\frac{a_{p_0}^{-1/2} r_p(\mathbf{x}) - \hat{r}_p(\mathbf{x}_i)}{h}\right] \right|,$$

where K is the kernel of the density estimate (here we use the Gaussian kernel). Now, using the mean value theorem and assuming $C_K = \sup_x |K'(x)| < \infty$ (i.e., bounded first derivative), we have

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - \hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \leq C_K \frac{\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{r}_p(\mathbf{x}) - a_{p_0}^{-1/2} r_p(\mathbf{x})|}{h^2}. \tag{6}$$

Note that under the assumption of l_p-symmetry, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ have $\sqrt{n}$ convergence (see Cator and Lopuhaä, 2011), and hence $\hat{r}_p(\mathbf{x})$ also has $\sqrt{n}$-convergence to $a_{p_0}^{-1/2} r_p(\mathbf{x})$. In fact, following arguments similar to Zuo (2003), we have $\sqrt{n}\sup_{\mathbf{x}\in A} |\hat{r}_p(\mathbf{x}) - a_{p_0}^{-1/2} r_p(\mathbf{x})| = O_p(1)$. So, if we have $h = O(n^{-1/4+\eta})$ for any $\eta > 0$ (which holds for the Sheather-Jones (1991) bandwidth), we get

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - \hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \xrightarrow{P} 0.$$

Now, consider the second term in (5) and note that
$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))|$

$$\leq \sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x})) - g_{p,h}^{*}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| + \sup_{\mathbf{x}\in\mathbb{R}^d} |g_{p,h}^{*}(a_{p_0}^{-1/2} r_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))|,$$

where $g_{p,h}^{*}(a_{p_0}^{-1/2} r_p(\mathbf{x})) = (nh)^{-1} \sum_{i=1}^{n} K[\{a_{p_0}^{-1/2} r_p(\mathbf{x}) - a_{p_0}^{-1/2} r_p(\mathbf{x}_i)\}/h]$. Using triangle inequality on the first term, we get
$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))|$

$$\leq C_K \frac{\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{r}_p(\mathbf{x}) - a_{p_0}^{-1/2} r_p(\mathbf{x})|}{h^2} + \sup_{\mathbf{x}\in\mathbb{R}^d} |g_{p,h}^{*}(a_{p_0}^{-1/2} r_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))|.$$

Now, from the uniform convergence properties of the kernel density estimate $g_{p,h}^{*}$ (see, e.g., Silverman, 1998) and arguments similar as above, we have

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(a_{p_0}^{-1/2} r_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \xrightarrow{P} 0. \qquad \square$$

Proof of Theorem 1 : Define $A_n = \{\mathbf{x} : r_p(\mathbf{x}) \in [\zeta_{1n}, \zeta_{2n}]\} \subset \mathbb{R}^d$. Note that using the mean value theorem on the logarithmic function, one gets

$$\sup_{\mathbf{x}\in A_n} |\ln \hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - \ln g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))|$$

$$\leq \sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \sup_{\mathbf{x}\in A_n} \left\{\frac{1}{\xi_n(\mathbf{x})}\right\},$$

where $\xi_n(\mathbf{x}) = \lambda_n \hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) + (1-\lambda_n) g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))$ for some $\lambda_n \in (0,1)$, and hence it converges to $g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))$ $[= a_{p_0}^{1/2} g_p(r_p(\mathbf{x}))]$ as $n \to \infty$ (follows from Lemma 3). Since the Sheather-Jones

bandwidth is of the order $O(n^{-1/5})$, from the proof of Lemma 3, it follows that $\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{g}_{p,h}(\hat{r}_p(\mathbf{x}))- g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| = O_P(n^{-1/10})$. So, under the condition $(C2)$, we have

$$\sup_{\mathbf{x}\in A_n} |\ln \hat{g}_{p,h}(\hat{r}_p(\mathbf{x})) - \ln g_p^{(a)}(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \xrightarrow{P} 0.$$

Following similar arguments as above, one can show that

$$\sup_{\mathbf{x}\in A_n} |\ln \hat{r}_p(\mathbf{x}) - \ln(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \leq \sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{r}_p(\mathbf{x}) - (a_{p_0}^{-1/2} r_p(\mathbf{x}))| \sup_{\mathbf{x}\in A_n} \left\{\frac{1}{\gamma_n(\mathbf{x})}\right\},$$

where $\gamma_n(\mathbf{x}) = \lambda_n \hat{r}_p(\mathbf{x}) + (1-\lambda_n) a_{p_0}^{1/2} r_p(\mathbf{x})$ for some $\lambda_n \in (0,1)$, and it converges to $a_{p_0}^{1/2} r_p(\mathbf{x})$. So, using $(C1)$ and $\sqrt{n}$ uniform in probability convergence of $\hat{r}_p(\mathbf{x})$, we have

$$\sup_{\mathbf{x}\in A_n} |\ln \hat{r}_p(\mathbf{x}) - \ln(a_{p_0}^{-1/2} r_p(\mathbf{x}))| \xrightarrow{P} 0, \text{and hence } \sup_{\mathbf{x}\in A_n} |\ln \hat{f}_n(\mathbf{x},p) - \ln(a_{p_0}^{d/2} f(\mathbf{x},p))| \xrightarrow{P} 0. \tag{7}$$

Now, define $I_n = \{i : \mathbf{X}_i \in A_n\}$ and note that using the triangle inequality, one gets

$$\left|\frac{1}{n}\sum_{i\ \in I_n} \ln \hat{f}_n(\mathbf{x}_i,p) - E_{p_0}\{\ln(a_{p_0}^{d/2} f(\mathbf{X},p))\}\right|$$

$$\leq \sup_{\mathbf{x}\in A_n} |\ln \hat{f}_n(\mathbf{x},p) - \ln(a_{p_0}^{d/2} f(\mathbf{x},p))| + \left|\frac{1}{n}\sum_{i\ \in I_n} \ln(a_{p_0}^{d/2} f(\mathbf{x}_i,p)) - E_{p_0}\{\ln(a_{p_0}^{d/2} f(\mathbf{X},p))\}\right|,$$

where $E_{p_0}\{.\}$ denotes the expectation with respect to $f(\mathbf{x},p_0)$. Convergence of the first part to 0 follows from (7). Also, from WLLN, we have

$$\left|\frac{1}{n}\sum_{i\ \in I_n} \ln \hat{f}_n(\mathbf{x}_i,p) - E_{p_0}\{\ln(a_{p_0}^{d/2} f(\mathbf{X},p)) I(\mathbf{X}\in A_n)\}\right| \xrightarrow{P} 0.$$

Now, $E_{p_0}\{\ln(a_{p_0}^{d/2} f(\mathbf{X},p)) I(\mathbf{X}\in A_n)\} \to E_{p_0}\{\ln(a_{p_0}^{d/2} f(\mathbf{X},p))\}$ as $n\to\infty$. So, for all $p \geq 1$, we have

$$\frac{1}{n}\sum_{i\ \in I_n} \ln \hat{f}_n(\mathbf{x}_i,p) \xrightarrow{P} E_{p_0}\{\ln(a_{p_0}^{d/2} f(\mathbf{X},p))\}.$$

Note that $E_{p_0}\{\ln(a_{p_0}^{d/2} f(\mathbf{X},p))\} = \ln(a_{p_0}^{d/2}) + E_{p_0}\{\ln f(\mathbf{X},p)\}$, and using Jensen's inequality on the logarithmic function, we have $E_{p_0}\{\ln f(\mathbf{X},p)\} < E_{p_0}\{\ln f(\mathbf{X},p_0)\}$ for all $p \neq p_0$. Also note that both $\hat{f}_n(\mathbf{x},p)$ and $E_{p_0}\{\ln f(\mathbf{X},p)\}$ are continuous functions of p. So, the proof now follows from the uniqueness conditions for the respective maximizers (see Lemma 5.10 in van der Vaart, 2000). □

Proof of Theorem 2 : The misclassification rate of the classifier d_1 is given by

$$\Delta(\mathsf{d}_1) = \frac{1}{J}\sum_{j=1}^{J} \int \theta_j(\mathbf{x},\hat{p}_n) f_j(\mathbf{x},p_0) d\mathbf{x},$$

where $\theta_j(\mathbf{x},\hat{p}_n) = P\{\hat{r}_{\hat{p}_n,j}(\mathbf{x}) > \hat{r}_{\hat{p}_n,i}(\mathbf{x})\}$ for some $1 \leq i \neq j \leq J$. From Theorem 1, it follows that $\hat{p}_n \xrightarrow{P} p_0$ and from the proof of Theorem 1, for any $\mathbf{x}\in\mathbb{R}^d$, $p \geq 1$, and $j = 1,2,\ldots,J$,

we have $\hat{r}_{\hat{p}_n,j}(\mathbf{x}) \xrightarrow{P} a_{p_0}^{-1/2} r_{p_0,j}(\mathbf{x})$, where $a_{p_0}^{-1/2}$ is a constant that depends on ψ, $\mathbf{\Sigma}$ and p_0. So, if $r_{p_0,j}(\mathbf{x}) < r_{p_0,i}(\mathbf{x})$ for all $i \neq j$ (or equivalently, $f_j(\mathbf{x}, p_0) > f_i(\mathbf{x}, p_0)$ for all $i \neq j$), $\theta_j(\mathbf{x}, p_0)$ converges to 0, otherwise it converges to 1. Therefore, by an application of the Dominated Convergence Theorem (DCT), we have the convergence of $\Delta(\mathsf{d}_1)$ to the Bayes risk. □

Lemma 4 : For any fixed $\mathbf{x} \in \mathbb{R}^d$, under $(C1)$ and $(C2)$, $\hat{f}(\mathbf{x}, \hat{p}_n) \xrightarrow{P} a_{p_0}^{d/2} f(\mathbf{x}, p_0)$ as $n \to \infty$.

Proof of Lemma 4 : Recall that $\hat{f}(\mathbf{x}, \hat{p}_n) = C_{\hat{p}_n,d} \frac{\hat{g}_{\hat{p}_n,h}(\hat{r}_{\hat{p}_n}(\mathbf{x}))}{\hat{r}_{\hat{p}_n}(\mathbf{x})^{d-1}}$, where $C_{p,d} = \frac{p^{d-1}\Gamma(d/p)}{2^d\{\Gamma(1/p)\}^d}$. Now, from the continuous mapping theorem and Theorem 1, we get

$$C_{\hat{p}_n,d} \xrightarrow{P} C_{p,d} \text{ and } \hat{r}_{\hat{p}_n}(\mathbf{x})^{d-1} \xrightarrow{P} \{a_{p_0}^{-1/2} r_{p_0}(\mathbf{x})\}^{d-1} \text{ for } d > 1. \tag{8}$$

Using the triangle inequality, we now have
$|\hat{g}_{\hat{p}_n,h}(\hat{r}_{\hat{p}_n}(\mathbf{x})) - g_{p_0}^{(a)}(a_{p_0}^{-1/2} r_{p_0}(\mathbf{x}))|$

$$\leq \sup_{r_{p_0}(\mathbf{x})} |\hat{g}_{\hat{p}_n,h}(r_{p_0}(\mathbf{x})) - g_{p_0}(r_{p_0}(\mathbf{x}))| + |g_{p_0}(\hat{r}_{\hat{p}_n}(\mathbf{x})) - g_{p_0}^{(a)}(a_{p_0}^{-1/2} r_{p_0}(\mathbf{x}))|.$$

Now, using arguments similar to proof of Theorem 1, and uniform convergence property of the kernel density estimate (see, e.g., Silverman, 1998), one can show that the first term goes to 0. From the continuity of the density function g_p, the second term also goes to 0. Combining both, we have $\hat{g}_{\hat{p}_n,h}(\hat{r}_{\hat{p}_n}(\mathbf{x})) \xrightarrow{P} g_{p_0}^{(a)}(a_{p_0}^{-1/2} r_{p_0}(\mathbf{x}))[= a_{p_0}^{1/2} g_{p_0}(r_{p_0}(\mathbf{x}))]$ as $n \to \infty$. For any fixed $\mathbf{x} \in \mathbb{R}^d$, the result follows from Slutsky's lemma and eqn (8). □

Proof of Theorem 3 : For simplicity, we consider the case when $J = 2$. Note that for $J > 2$, we use pairwise classification, and hence one can get the proof by repeating the same argument for each of the $\binom{J}{2}$ pair of classes. Note that using Lemma 4 for the two class problem, we get

$$(\hat{f}_{h_1}(\mathbf{x}, \hat{p}_{1n_1}), \hat{f}_{h_2}(\mathbf{x}, \hat{p}_{2n_2})) \xrightarrow{P} (a_{1p_1}^{d/2} f(\mathbf{x}, p_1), a_{2p_2}^{d/2} f(\mathbf{x}, p_2)), \text{ where } a_{1p_1}, a_{2p_2} > 0.$$

The leave-one-out cross-validation estimate of the error rate is given by

$$\Delta_{\mathbf{n}}^{CV}(k) = \sum_{i=1, j\neq i}^{2} \frac{\pi_i}{n_i} \sum_{l=1}^{n_i} I\left\{ \frac{\hat{f}_{h_i}^{-il}(\mathbf{x}_{il}, \hat{p}_{in_i})}{\hat{f}_{h_j}^{-il}(\mathbf{x}_{jl}, \hat{p}_{jn_j})} \geq k_i \right\}.$$

where $\mathbf{n} = (n_1, n_2)$, $k_1 = 1/k$, $k_2 = k$, and $\hat{f}_{h_j}^{-il}$ denotes the leave-one-out density estimate, where $\mathbf{x}_{il}$ is not used as a training data point. Also, define

$$\Delta(k) = \sum_{i=1, j\neq i}^{2} \pi_i P\left\{ \frac{a_{ip_i}^{d/2} f(\mathbf{X}, p_i)}{a_{jp_j}^{d/2} f(\mathbf{X}, p_j)} \geq k_i \;\middle|\; \mathbf{X} \in i\text{-th class} \right\},$$

$k_{\mathbf{n}} = \arg\min_k \Delta_{\mathbf{n}}^{CV}(k)$ and $k_0 = \arg\min_k \Delta(k)$. Note that $\Delta(k)$ possesses a unique minima k_0 $(= a_{1p_1}^{d/2}\pi_2 / a_{2p_2}^{d/2}\pi_1)$, and $\Delta(k_0)$ denotes the error rate of the Bayes classifier d_B. Now from the consistency of the density estimates, and using the same arguments as in Lemma 3 of Dutta and Ghosh (2011), one can show that $\sup_k |\Delta_{\mathbf{n}}^{CV}(k) - \Delta(k)| \xrightarrow{P} 0$, this implies that $k_{\mathbf{n}} \xrightarrow{P} k_0$ as

$\min\{n_1, n_2\} \to \infty$. Now, using the continuous mapping theorem and continuity of the underlying densities, for any fixed $\mathbf{x} \in \mathbb{R}^d$, we have $\mathsf{d}_2(\mathbf{x}) \xrightarrow{P} \mathsf{d}_B(\mathbf{x})$ and hence $I(\mathsf{d}_2(\mathbf{x}) \neq i) \xrightarrow{P} I(\mathsf{d}_B(\mathbf{x}) \neq i)$ for $i = 1, 2$. This implies that

$$\left| \sum_{j=1}^{J} \pi_j P(\mathsf{d}_2(\mathbf{X}) \neq j \mid \mathbf{X} \in j\text{-th class}) - \sum_{j=1}^{J} \pi_j P(\mathsf{d}_B(\mathbf{X}) \neq j \mid \mathbf{X} \in j\text{-th class}) \right|$$

$$\leq \sum_{j=1}^{J} \pi_j \left| P(\mathsf{d}_2(\mathbf{X}) \neq j \mid \mathbf{X} \in j\text{-th class}) - P(\mathsf{d}_B(\mathbf{X}) \neq j \mid \mathbf{X} \in j\text{-th class}) \right| \to 0,$$

as $\min\{n_1, n_2\}$ tends to infinity (follows from the Dominated Convergence Theorem). □

References

[1] Cator, E. A. and Lopuhaä, H. P. (2011) Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *To appear in Bernoulli.*

[2] Chakraborty, B. and Chaudhuri, P. (1996) On a transformation and re-transformation technique for constructing an affine equivariant multivariate median. *Proc. Amer. Math. Soc.*, **124**, 1529-1537.

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

[4] Christmann, A., Fischer, P. and Joachims, T. (2002) Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassification. *Comput. Statist.*, **17**, 273-287.

[5] Cuevas, A. and Fraiman, R. (2009) On depth measures and dual statistics. A methodology for dealing with general data. *J. Multivariate Anal.*, **100**, 753-766.

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

[7] Dutta, S. and Ghosh, A. K. (2011) On robust classification using projection depth. *To appear in Ann. Inst. Stat. Math.*

[8] Dutta, S., Ghosh, A. K. and Chaudhuri, P. (2011) Some intriguing properties of Tukey's half-space depth. *To appear in Bernoulli.*

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

[10] Fraiman, R., Liu, R. Y. and Meloche, J. (1997) Multivariate density estimation by probing depth. In L_1-*Statistical Procedures and Related Topics. IMS Lecture Notes* (ed. Y. Dodge), Inst. Math. Stat., Hayward, California, **31**, 415-430.

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

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

[13] Ghosh, A. K. and Hall, P. (2008) On error rate estimation in nonparametric classification. *Statistica Sinica*, **18**, 1081-1100.

[14] Hartikainen, A. and Oja, H. (2006) On some parametric, nonparametric and semiparametric discrimination rules. *DIMACS Ser. Math. Theo. Comput. Sci.*, (R. Liu, R. Serfling and D. L. Souvaine ed.), **72**, 61-70.

[15] Hastie, T., Tibshirani, R. and Friedman, J. (2009) *Elements of Statistical Learning Theory.* Wiley, New York.

[16] Hoberg, R. and Mosler, K. (2006) Data analysis and classification with the zonoid depth. *DIMACS Ser. Math. Theo. Comput. Sci.*, (R. Liu, R. Serfling and D. L. Souvaine ed.), **72**, 49-59.

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

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

[19] Jornsten, R. (2004) Clustering and classification based on the L_1 data depth. *J. Multivariate Anal.*, **90**, 67-89.

[20] Li, J., Cuesta-Albertos, J. A. and Liu, R. (2011) Nonparametric classification procedures based on DD-plot. *Available at http://personales.unican.es/cuestaj/DDPlot_ Classification.pdf.*

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

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

[23] López-Pintado, S. and Romo, J. (2006) Depth-based classification for functional data. *DIMACS Ser. Math. Theo. Comput. Sci.*, (R. Liu, R. Serfling and D. L. Souvaine ed.), **72**, 103-119.

[24] Rousseeuw, P.J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. Wiley, New York.

[25] Sheather, S. J. and Jones, M. C. (1991) A reliable data based bandwidth selection method for kernel density estimation. *J. Royal Statist. Soc. Ser. B*, **53**, 683-690.

[26] Silverman, B. W. (1998) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

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

[28] van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge Univ. Press, London.

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

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

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

[32] Zuo, Y. (2003) Projection-based depth functions and associated medians. *Ann. Statist.*, **21**, 1460-1490.