DD-Classifier: Nonparametric Classification Procedure Based on *DD*-plot[1]

Jun Li[2], Juan A. Cuesta-Albertos[3], Regina Y. Liu[4]

Abstract

Using the *DD*-plot (depth-versus-depth plot), we introduce a new nonparametric classification algorithm and call it a *DD*-classifier. The algorithm is completely nonparametric, and requires no prior knowledge of the underlying distributions or of the form of the separating curve. Thus it can be applied to a wide range of classification problems. The algorithm is completely data driven and its classification outcome can be easily visualized on a two-dimensional plot regardless of the dimension of the data. Moreover, it is easy to implement since it bypasses the task of estimating underlying parameters such as means and scales, which is often required by the existing classification procedures. We study the asymptotic properties of the proposed *DD*-classifier and its misclassification rate. Specifically, we show that it is asymptotically equivalent to the Bayes rule under suitable conditions. The performance of the classifier is also examined by using simulated and real data sets. Overall, the proposed classifier performs well across a broad range of settings, and compares favorably with existing classifiers. Finally, it can also be robust against outliers or contamination.

Key words: Classification, data depth, *DD*-plot, *DD*-classifier, maximum depth classifier, misclassification rates, nonparametric, robustness.

1 Introduction

Classification is one of the most practical subjects in statistics. It has many important applications in different fields, such as disease diagnosis in medical sciences, flaw detection

[1]Jun Li, Assistant Professor, Department of Statistics, University of California, Riverside, CA 92521 (Email: *jun.li@ucr.edu*). Juan A. Cuesta-Albertos, Professor, Department of Mathematics, University of Cantabria, 39071 Santander, Spain (Email: *juan.cuesta@unican.es*). Regina Y. Liu, Professor, Department of Statistics & Biostatistics, Rutgers University, Hill Center, Piscataway, NJ 08854-8019 (E-mail: *rliu@stat.rutgers.edu*). The authors thank Bing Li and Anil K. Ghosh for data sharing.

[2]Research in part supported by National Science Foundation Grant DMS-0907655

[3]Research in part supported by the Spanish Ministerio de Ciencia y Tecnología, grants MTM2008-0607-C02-02 and PR2009-0355

[4]Research in part supported by National Science Foundation Grant DMS-0707053

in engineering, and risk identification in finance, to name a few. Many existing classification algorithms assume either certain parametric distributions for the data or certain forms of separating curves or surfaces. These parametric classifiers are suboptimal and of limited use in practical applications where little information about the underlying distributions is available a priori. In comparison, nonparametric classifiers are usually more flexible in accommodating different data structures, and are hence more desirable. In this paper, we propose and study a new nonparametric classifier using the *DD*-plot (depth vs depth plot) introduced in Liu et al. (1999). We shall refer to this classifier as a *DD*-classifier. Roughly speaking, for two given multivariate samples, its *DD*-plot represents the depth values of those sample points with respect to the two underlying distributions, and thus transforms the samples in any dimension to a simple two-dimensional scatter plot. The idea of our *DD*-classifier is to look for a curve that best separates the two samples in their *DD*-plot, in the sense that the separation yields the smallest classification error in the *DD*-plot. Clearly, the best separating curve in the *DD*-plot leads to a classification rule in the original sample space of the two samples. Some obvious advantages of this approach are: 1) The best separating curve in the *DD*-plot is determined automatically by the underlying probabilistic geometry of the data and is completely data driven. Therefore, the *DD*-classifier approach is fully nonparametric. 2) Since the depth transformation has the standardization effect on the data, the *DD*-classifier bypasses the task of estimating parameters such as means and scales, which is often required by the existing classification procedures. 3) The classification outcome can be easily visualized in the two-dimensional *DD*-plot. This is a much simpler task than tracking the classification outcome in the original sample space, which can be forbiddingly difficult if the samples are of high dimensions.

In the last two decades, data depth has emerged as a powerful analysis tool in various areas of multivariate statistics, since it can characterize the centrality of a distribution as well as motivate nonparametric robust statistical methodologies. In fact, it has already offered several promising solutions to classification problems. For instance, Christmann and Rousseeuw (2001) and Christmann, et al. (2002) applied the idea of regression depth (see Rousseeuw and Huber, 1999) to classification. Ghosh and Chaudhuri (2005a) used half-space depth and regression depth to construct linear and nonlinear separating curves or surfaces. In those depth based methods, a finite dimensional parametric form (usually linear or quadratic) for the separating surface is often assumed. Thus, these classifiers are not fully nonparametric. The possibility of using the maximum depth for nonparametric classification was raised in Liu (1990). This was carried out fully by Ghosh and Chaudhuri (2005b), who developed the notion of maximum depth classifier into a full-fledged nonparametric classification rule.

This classification rule assigns the observation to the group for which it attains the highest depth value, since higher depth values should correspond to more central positions within the group. This classification rule is intuitively appealing and fully nonparametric, but it performs well only when the populations differ in location only and the prior probabilities of the populations are equal. Ghosh and Chaudhuri (2005b) recognized this limitation and proposed a modified classification rule. However, the modified approach is valid only for elliptical distributions when the half-space depth is used. This approach requires estimating several unknown parameters, some of which involve complicated estimation techniques. Therefore, the modified approach is complicated and no longer fully nonparametric, which diminishes its appeal to practitioners. Recently, Cui et al. (2008) considered a maximum depth classifier based on a modified projection depth. However, this classifier appears to work well only under normal settings.

In this paper, we show that our proposed *DD*-classifier is asymptotically equivalent to Bayes rule under elliptical distributions. Furthermore, we show that it performs well in a broad range of settings, including non-elliptical distributions or distributions differing in scale. The latter is a case where the maximum depth classifier is known to fail. Our comparison studies show that the *DD*-classifier often outperforms the maximum depth classifier, and is comparable or better than the k-nearest neighbor method. Besides the advantages that we had stressed earlier of being nonparametric, completely data-driven, and simple to visualize, the proposed *DD*-classifier is also easy to implement and robust against outliers and extreme values.

The rest of this paper is organized as follows. In Section 2, we provide a brief review of data depth, *DD*-plot and notations. In Section 3, we describe in detail the proposed *DD*-classifier, and argue heuristically why it should yield the best separating curve for two competing classes. In Section 4, we study the asymptotic properties of the *DD*-classifier and its misclassification rate. In particular, we show that the *DD*-classifier is asymptotically equivalent to the Bayes rule under suitable conditions. In Section 5, we address some practical issues regarding the implementation of *DD*-classifiers. In Section 6, we conduct several simulation studies to evaluate the performance of the *DD*-classifier. In Section 7, we demonstrate some applications of our classifier to real data sets. Concluding remarks are in Section 8. All the proofs are deferred to the Appendix.

2 Background Material on Data Depth and *DD*-plot

A data depth is a measure of "depth" or "centrality" of a given point with respect to a multivariate data cloud or its underlying distribution. For example, considering $\{X_1, \cdots, X_m\}$, a random sample from the distribution $F(\cdot)$ in $\mathbb{R}^d$ $(d \geq 1)$, the simplicial depth (Liu (1990)) of x w.r.t. F is defined as $SD_F(x) = P_F\{x \in s[X_1, \cdots, X_{d+1}]\}$, where $s[X_1, \cdots, X_{d+1}]$ is a closed simplex formed by $(d+1)$ random observations from F, and its sample version, $D_{F_m}(x)$, is given by,

$$SD_{F_m}(x) = \binom{m}{d+1}^{-1} \sum_{(*)} I(x \in s[X_{i_1}, \cdots, X_{i_{d+1}}]),$$

where (*) runs over all possible subsets of $\{X_1, \cdots, X_m\}$ of size $(d+1)$. A larger value of $D_{F_m}(x)$ indicates that x is contained in more simplices generated from the sample, and thus it lies deeper within the data cloud. Hence, the simplicial depth is a measure of centrality of a given point with respect to a multivariate data cloud or its underlying distribution.

There are many other notions of data depth (see, e.g., Liu et al. (1999), Zuo and Serfling (2000)). In this paper, we use Mahalanobis depth to cover the well studied Gaussian case, and half-space depth, simplicial depth, and projection depth to explore the robustness aspect of our approach. The last three depths are geometric and thus completely nonparametric. We review the definitions of Mahalanobis depth, half-space depth and projection depth.

Definition 2.1. The *Mahalanobis depth* (Mahalanobis (1936)) at x w.r.t. F is defined as

$$MhD_F(x) = [1 + (x - \mu_F)'\Sigma_F^{-1}(x - \mu_F)]^{-1},$$

where μ_F and Σ_F are the mean vector and covariance matrix of F respectively. The sample Mahalanobis depth is obtained by replacing μ_F and Σ_F with their sample estimates.

Definition 2.2. The *half-space depth* (Hodges (1955), Tukey (1975)) at x w.r.t. F is defined as $HD_F(x) = \inf_H\{P_F(H) : H \text{ is a closed half-space in } \mathbb{R}^d \text{ and } x \in H\}$. Its sample version is $HD_{F_m}(x)$, obtained by replacing F in $HD_F(x)$ by the empirical distribution F_m.

Definition 2.3. The *projection depth* (Stahel (1981), Donoho (1982), Donoho and Gasko (1992), Zuo (2003)) at x w.r.t. F is defined as

$$PD_F(x) = \left[1 + \sup_{\|u\|=1} |u'x - \mathrm{Med}(F_u)| \,/\mathrm{MAD}(F_u)\right]^{-1},$$

where F_u is the distribution of $u'x$, $\mathrm{Med}(F_u)$ is the median of F_u, and $\mathrm{MAD}(F_u)$ is the median absolute deviation of F_u. The sample version of $PD_F(x)$ is $PD_{F_m}(x)$, obtained by replacing the median and MAD with their sample estimates.

For convenience, we use the notation $D(\cdot)$ to express any valid notion of depth, unless a particular notion is to be singled out.

Next we briefly review the so-called *DD*-plot (depth vs depth plot). Let $\{X_1, ..., X_m\}(\equiv \mathbf{X})$ and $\{Y_1, ..., Y_n\}(\equiv \mathbf{Y})$ be two random samples respectively from distributions F and G which are defined on $\mathbb{R}^d$. The *DD*-plot is defined as

$$DD(F, G) = \{(D_F(x), D_G(x)), x \in \mathbf{X} \cup \mathbf{Y}\}.$$

If both F and G are unknown, the DD-plot is then defined as

$$DD(F_m, G_n) = \{(D_{F_m}(x), D_{G_n}(x)), x \in \mathbf{X} \cup \mathbf{Y}\} \tag{2.1}$$

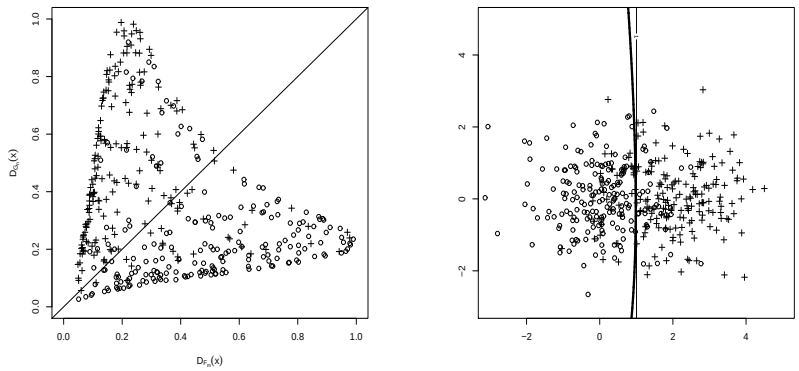
The *DD*-plot was first introduced by Liu et al. (1999) for graphical comparisons of two multivariate distributions or samples based on data depth. It is always a two dimensional graph regardless of the dimensions of the samples. It was shown in Liu et at. (1999) that different distributional differences, such as location, scale, skewness or kurtosis differences, are associated with different graphic patterns in the *DD*-plot. Therefore, *DD*-plots can provide simple diagnostic tools for visual comparisons of two samples of any dimension. For example, Li and Liu (2004) derives several rigorous nonparametric tests of multivariate locations and scales by detecting possible departures from the expected patterns of graphs in *DD*-plots. In this paper, we shall show that *DD*-plots can also detect other differences between two populations, based on which a novel classifier can be constructed.

3 *DD*-classifier

For simplicity, we consider only two-class classification problem in this paper, although the proposed classification approach can easily extend to multi-class problems by incorporating the method of majority voting (see, e.g., Friedman (1996)). Again let $\{X_1, ..., X_m\}(\equiv \mathbf{X})$ and $\{Y_1, ..., Y_n\}(\equiv \mathbf{Y})$ be two random samples from respectively F and G, which are distributions defined on $\mathbb{R}^d$. From the definition of DD-plot in (2.1). We see that if $F = G$, the *DD*-plot should be concentrated along the 45 degree line. If the two distributions F and G differ, the *DD*-plot would exhibit a noticeable departure from the 45 degree line. This is shown in Figure 1(a), where the *DD*-plot is constructed for two bivariate normal samples: one is from the standard bivariate normal distribution while the other is with a mean shift to $(2, 0)'$. Both sample sizes are 200. The plot is constructed using the Mahalanobis depth.

To facilitate the identification of sample points from one sample versus the other, we use different symbols to indicate the membership of the sample points. For example, in Figure 1(a), the "o"s represent the observations from $\mathbf{X}$, and the "+"s represent those from $\mathbf{Y}$.

The *DD*-plot in Figure 1(a) shows quite clearly that the observations from the two different samples are now displayed around the 45 degree line in an almost symmetric manner. If we are to separate the two samples in the *DD*-plot using a line, the 45 degree line appears to be the best choice. In fact, if we use the 45 degree line as the separating line, its corresponding classification rule would assign x to F if $D_{F_m}(x) > D_{G_n}(x)$ and assign x to G otherwise. Note that this is actually the same as the maximum depth classifier studied in Ghosh and Chaudhuri (2005b). It was shown in Ghosh and Chaudhuri (2005b) that the maximum depth classifier is asymptotically equivalent to the Bayes rule if the two distributions have the same prior probabilities and are elliptical with only a location difference. Therefore, in this case, the best separating line between the two samples in the *DD*-plot should yield something very close to the best separating line between two samples in the original sample space $\mathbb{R}^2$. We can see this in Figure 1(b), which shows our samples in their original sample space $\mathbb{R}^2$. Again the "o"s represent the observations from $\mathbf{X}$, and the "+"s represent those from $\mathbf{Y}$. The thick curve is obtained by mapping the 45 degree line in the *DD*-plot back to the $\mathbb{R}^2$ space. The thin line is generated from the Bayes rule. Both curves are indeed very close.

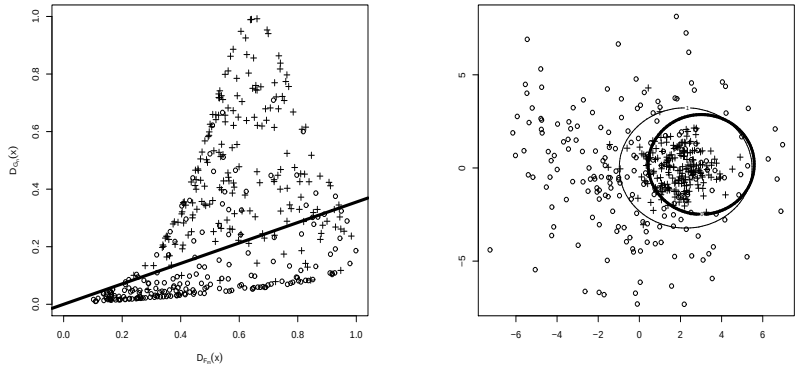


(a) *DD*-plot (b) scatter plot

Figure 1: Two bivariate normal samples with only a location difference

To consider scale differences, we multiply all the observations in $\mathbf{X}$ by 3, and view the new sample $\mathbf{X}$ as drawn from a bivariate normal distribution with mean $(0, 0)'$ and covariance matrix $9I_2$, where I_2 is the two dimensional identity matrix. Figure 2(a) shows the *DD*-plot of $\mathbf{Y}$ and the new sample $\mathbf{X}$. Compared with Figure 1(a), the two samples are no longer displayed symmetrically. The observations from $\mathbf{X}$ now seem to move towards the x-axis and the observations from $\mathbf{Y}$ towards the vertical line $x = 1$. For these two samples, if we

still use the maximum depth classifier, which is equivalent to drawing the 45 degree line and assigning the observations above the line to G and the ones below to F, it is obvious that this classification rule would assign most of the observations from $\mathbf{Y}$ to F, and thus yield a large misclassification rate. Therefore, this *DD*-plot clearly illustrates why the maximum depth classifier does not perform well when the distributions have different dispersion structures.



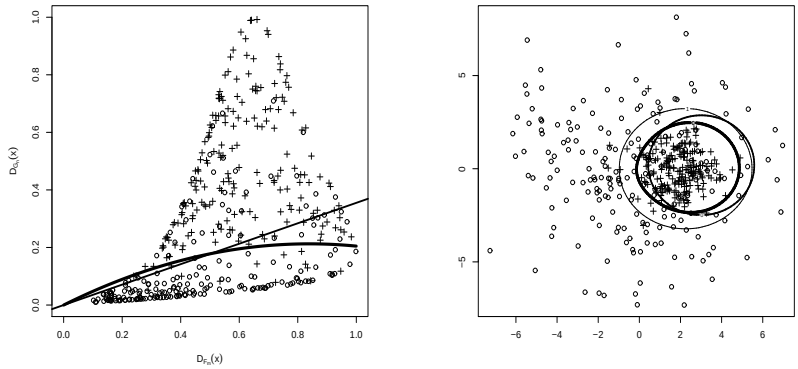
(a) *DD*-plot (b) scatter plot

Figure 2: Two bivariate normal samples with location and scale differences *with line separation in DD-plot*

Although the two samples in the *DD*-plot in Figure 2(a) no longer scatter in a symmetric pattern, it is still visible that there exists a line which can well separate the two samples, such as the one drawn in Figure 2(a). Thus the *DD*-plot remains a useful tool to visualize the separation of two samples. Further investigation shows that classification error can be reduced if the separating line in Figure 2(a) is replaced with a suitable polynomial, as seen in Figure 3(a). The thick solid circles in Figures 2(b) and 3(b) are the separating curves in the original sample space corresponding respectively to the line in 2(a) and the polynomial in 3(a). The largest circle in Figure 2(b) is the separating curve derived from the Bayes rule.

The observations from Figures 1 to 3 seem to suggest the general phenomenon that the curve (or line) best separating the two samples in the *DD*-plot will also yield the best separating curve between two samples in the original sample space. This phenomenon is confirmed in the following proposition for uni-modal and elliptical distributions.

Proposition 1 *Let $f_1(\cdot)$ and $f_2(\cdot)$ be the density functions of F and G respectively. Assume that they are both from the elliptical family and have the following form,*



(a) *DD*-plot (b) scatter plot

Figure 3: Two bivariate normal samples with location and scale differences *with polynomial separation in DD-plot*

$$f_i(x) = c_i|\Sigma_i|^{-1/2}h_i\left((x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right) \quad i = 1, 2, \tag{3.1}$$

where the $h_i(\cdot)$*'s are strictly decreasing functions. If* D_F *and* D_G *are strictly increasing functions of* f_1 *and* f_2 *respectively, then the Bayes rule is equivalent to*

$$\begin{cases} D_G(x) > r(D_F(x)) & \Longrightarrow \textit{assign } x \textit{ to } G \\ D_G(x) < r(D_F(x)) & \Longrightarrow \textit{assign } x \textit{ to } F \end{cases}$$

where $r(\cdot)$ *is some real increasing function.*

We note that the Mahalanobis depth, the half-space depth, the simplicial depth, and the projection depth all satisfy the condition required of depths above. (See, Liu (1990), Arcones et al. (1994), Zuo and Serfling (2000), Zuo (2003).)

Proposition 1 implies that for elliptical distributions, the best separating curve between two samples in the $\mathbb{R}^d$ space is equivalent to a curve best separating the two samples in the *DD*-plot. For non-elliptical distributions, we expect that the best separating function in the *DD*-plot would also lead to a reasonably good classifier since data depth characterizes the underlying distributions by centrality. Note that this separating function in the *DD*-plot may not be increasing for non-elliptical settings. In this case, we may consider obtaining also the best separating function from the *DD*-plot with the axes for D_G and D_F interchanged, namely $DD(G, F)$. This best separating function will give a classifier which is different from

the one derived from $DD(F, G)$. In principle, we can choose the one that yields the smaller misclassification rate as the recommended classifier. This would simply repeat the same procedure with the roles of D_F and D_G interchanged. Thus, for the rest of the paper, we illustrate our approach using only the $DD(F, G)$-plot. Note that the classification results can only improve if we do consider both *DD*-plots.

To find the function $r(\cdot)$ which separates best the two samples in the *DD*-plot, we may restrict ourselves to polynomials, since in principle any smooth function can be well approximated by a polynomial with suitable degree.

Note that the separating polynomial corresponding to a reasonable classification rule should pass through the origin in the *DD*-plot. To see this, consider the observations corresponding to $(0, 0)$ in the *DD*-plot. Since they have zero depth values with respect to both samples, their memberships are not clear. Hence, a reasonable classification rule would place those observations on the separating curve, indicating that they can be from either sample. This point is also borne out in the setting of Proposition 1, whose proof clearly shows that $r(\cdot)$ satisfies $r(0) = 0$. Thus, to search for the polynomial which separates best the two samples in the *DD*-plot, we consider polynomials of the form $r_{\mathbf{a}}(x) = \sum_{i=1}^{k_0} a_i x^i$. Here k_0 is the degree of the polynomial and is a predetermined known integer, and $\mathbf{a} = (a_1, \ldots, a_{k_0}) \in \mathbb{R}^{k_0}$ is the coefficient vector of the polynomial. Now we consider the following classification algorithm

$$\begin{cases} D_{G_n}(x) > r_{\mathbf{a}}(D_{F_m}(x)) & \Longrightarrow \text{assign } x \text{ to } G \\ D_{G_n}(x) < r_{\mathbf{a}}(D_{F_m}(x)) & \Longrightarrow \text{assign } x \text{ to } F, \end{cases}$$

and our goal is, for a prefixed k_0, to find the optimal $\mathbf{a}$ which minimizes the overall misclassification rate. We denote this optimal $\mathbf{a}$, if exists, by $\mathbf{a}_0$ and refer to the corresponding classifier as the *DD*-classifier from now on.

For any given $\mathbf{a} \in \mathbb{R}^{k_0}$, we can draw a polynomial corresponding to $D_{G_n}(x) = r_{\mathbf{a}}(D_{F_m}(x))$ in the *DD*-plot, assign the observations above the curve to G and those below it to F, and then calculate the empirical misclassification rate, namely

$$\hat{\Delta}_N(\mathbf{a}) = \frac{\pi_1}{m} \sum_{i=1}^{m} I_{\{D_{G_n}(X_i) > r_{\mathbf{a}}(D_{F_m}(X_i))\}} + \frac{\pi_2}{n} \sum_{i=1}^{n} I_{\{D_{G_n}(Y_i) < r_{\mathbf{a}}(D_{F_m}(Y_i))\}}. \tag{3.2}$$

Here the π_i are the prior probabilities of the two classes, $N = (m, n)$, $I_{\{A\}}$ is the indicator function which takes 1 if A is true and 0 otherwise. Therefore, we propose to estimate the optimal $\mathbf{a}_0$ by $\hat{\mathbf{a}}_N$, which minimizes the empirical misclassification rate $\hat{\Delta}_N(\mathbf{a})$ in (3.2). More

specifically, if $\hat{\mathbf{a}}_N = \text{argmin}\{\hat{\Delta}_N(\mathbf{a})\}$, our *DD*-classifier is

$$\begin{cases} D_{G_n}(x) > r_{\hat{\mathbf{a}}_N}(D_{F_m}(x)) & \Longrightarrow \text{assign } x \text{ to } G \\ D_{G_n}(x) < r_{\hat{\mathbf{a}}_N}(D_{F_m}(x)) & \Longrightarrow \text{assign } x \text{ to } F. \end{cases} \tag{3.3}$$

Applying this to the dataset in Figure 2 by following the steps described in Secion 5, we obtain a separating polynomial of degree 2 which minimizes the overall empirical misclassification rate. This polynomial is plotted in Figure 3(a) and its corresponding separating curve in the original sample space $\mathbb{R}^2$ is drawn as the thickest solid circle in Figure 3(b). The thinly dotted circle in Figure 3(b) is the separating curve obtained from the Bayes rule.

Figures 1(b) to 3(b) also show that, although we always focus on the separating polynomial in the *DD*-plot, when the polynomial is mapped back to the original sample space, the separating curve can be of any shape, depending on the structure of the data. Therefore, unlike many existing classification methods which require the pre-specification of certain parametric forms of the separating curves in the sample space, the form of the separating curves in the sample space obtained by our *DD*-classifier is automatically determined by the geometric structure of the data that underlie the *DD*-plot.

4 Properties of *DD*-classifier

In this section, we show the consistency of *DD*-classifier, and, also, we show that, under proper conditions, the proposed classification rule is asymptotically equivalent to the Bayes rule. Before stating the main results, we first introduce some definitions and notation. Since the depths we consider in this paper are bounded, without loss of generality, we assume that they are bounded by 1. Given $\mathbf{a} \in \mathbb{R}^{k_0}$, and $d_1, d_2 \in [0, 1]$, we define

$$C_{\mathbf{a}}(d_1, d_2) = \begin{cases} 1 & \text{if } d_2 > r_{\mathbf{a}}(d_1) \\ 0 & \text{if } d_2 \leq r_{\mathbf{a}}(d_1). \end{cases}$$

Our proposed *DD*-classifier and its empirical version in (3.3) can be represented respectively by $C_{\mathbf{a}}(D_F(x), D_G(x))$ and $C_{\hat{\mathbf{a}}_N}(D_{F_m}(x), D_{G_n}(x))$. Their associated misclassification rates are

$$\Delta(C_{\mathbf{a}}) = \pi_1 P_F\{z : C_{\mathbf{a}}(D_F(z), D_G(z)) = 1\} + \pi_2 P_G\{z : C_{\mathbf{a}}(D_F(z), D_G(z)) = 0\},$$

and

$$\Delta_N(\hat{C}_N) = \pi_1 P_F\{z : C_{\hat{\mathbf{a}}_N}(D_{F_m}(z), D_{G_n}(z)) = 1\} + \pi_2 P_G\{z : C_{\hat{\mathbf{a}}_N}(D_{F_m}(z), D_{G_n}(z)) = 0\}.$$

Here $\Delta_N(\hat{C}_N)$ can be viewed as the conditional misclassification probability given the training sample when the values $D_{G_n}(Z)$ and $r_{\hat{\mathbf{a}}_N}(D_{F_m}(Z))$ are used to classify the future observation Z. Its expectation $E\left(\Delta_N(\hat{C}_N)\right)$ w.r.t. the probability distribution of the training sample is the unconditional misclassification probability of the proposed DD-classifier.

Note that the family $\{C_{\mathbf{a}} : \mathbf{a} \in \mathbb{R}^{k_0}\}$ is not closed with respect to pointwise convergence, since its closure includes indicators of the union of at most k_0 disjoint intervals. For the latter case, we can use s to represent an indicator of a finite union of intervals, and consider the function

$$C_s(d_1, d_2) = s(d_1).$$

When this is applied to $(D_F(z), D_G(z))$, it classifies the point z based solely on its depth with respect to F. The misclassification error associated with this function s is

$$\Delta(C_s) = \pi_1 P_F\{z : C_s(D_F(z), D_G(z)) = 1\} + \pi_2 P_G\{z : C_s(D_F(z), D_G(z)) = 0\}.$$

We denote by Γ the family of *classification rules* composed of functions $C_{\mathbf{a}}, \mathbf{a} \in \mathbb{R}^{k_0}$, and of functions which are indicators of the union of at most k_0 disjoint intervals in $[0, 1]$.

Several lemmas are needed for proving the main result. In these lemmas we need to assume that, for any $\mathbf{a} \in \mathbb{R}^{k_0}$ and $\delta \in \mathbb{R}$,

$$P_F\{z : D_G(z) = r_{\mathbf{a}}(D_F(z))\} = P_G\{z : D_G(z) = r_{\mathbf{a}}(D_F(z))\} = 0 \tag{4.1}$$

$$P_F\{z : D_F(z) = \delta\} = P_G\{z : D_F(z) = \delta\} = 0. \tag{4.2}$$

Lemma 2 *If F and G satisfy (4.1) and (4.2), then there exists $C_0 \in \Gamma$ such that $\Delta(C_0) = \inf_{C\in\Gamma} \Delta(C)$.*

Remark 3 From the proof of Lemma 2, we can infer that the optimal classification rule corresponds to a fixed $\mathbf{a} \in \mathbb{R}^{k_0}$ or its limit. If the latter occurs in the setting that F and G belong to the elliptical family, then the optimal classification rule simply corresponds to an indicator function of one interval regardless the value k_0. This holds because r is increasing in this case.

Lemma 4 *Suppose that F and G satisfies (4.1) and (4.2). Furthermore, assume that both $D_F(\cdot)$ and $D_G(\cdot)$ are continuous and satisfy, as $\min(m, n) \to \infty$,*

$$\sup_x |D_{F_m}(x) - D_F(x)| \xrightarrow{a.s.} 0, \quad \sup_x |D_{G_n}(x) - D_G(x)| \xrightarrow{a.s.} 0. \tag{4.3}$$

Then we have, for every $C \in \Gamma$, as $\min(m, n) \to \infty$,

$$\hat{\Delta}_N(C) \to \Delta(C) \text{ a.s.}$$

The four data depths studied in this paper satisfy the conditions in Lemma 4, see, e.g. Liu (1990), Dümbgen (1992), Zuo and Serfling (2000) and Zuo (2003).

The following theorem shows that, under suitable conditions, $\hat{C}_N$, which minimizes the empirical misclassification rate, converges asymptotically to C_0, which minimizes the population misclassification rate.

Theorem 5 *Suppose that the assumptions in Lemma 4 hold. Let* $\hat{C}_N = \text{argmin}_{C\in\Gamma}\{\hat{\Delta}_N(C)\}$ *and* $C_0 = \text{argmin}_{C\in\Gamma}\{\Delta(C)\}$. *If* C_0 *is unique, we have, as* $\min(m,n) \to \infty$,

$$\hat{C}_N \to C_0 \ a.s.$$

where this a.s. convergence should be understood in the sense that

$$\begin{aligned} P_F\{z : \hat{C}_N(D_F(z), D_G(z)) \to C_0(D_F(z), D_G(z))\} &= 1 \\ P_G\{z : \hat{C}_N(D_F(z), D_G(z)) \to C_0(D_F(z), D_G(z))\} &= 1. \end{aligned}$$

We are now ready to state our main result.

Theorem 6 *Assume that the density functions* $f_1(\cdot)$ *and* $f_2(\cdot)$ *of* F *and* G *respectively are of the form in (3.1), with* $h_1(\cdot) = h_2(\cdot)$ *and* $\Sigma_1 = \Sigma_2$. *Assume that* $\pi_1 = \pi_2$. *If the depth function used in the classification algorithm is Mahalanobis depth, half-space depth, simplicial depth, or projection depth, then, as* $\min(m,n) \to \infty$, *if* $\mathbf{a}_1 = (1, 0, \ldots, 0)$, *we have*

$$E\left(\Delta_N(\hat{C}_N)\right) \longrightarrow \Delta(C_{\mathbf{a}_1}).$$

Remark 7 Following Proposition 1, it is easy to see that, under the assumptions of Theorem 6, the Bayes rule is equivalent to

$$\begin{cases} D_G(x) > D_F(x) & \Longrightarrow \text{assign } x \text{ to } G \\ D_G(x) < D_F(x) & \Longrightarrow \text{assign } x \text{ to } F. \end{cases}$$

Thus, $\Delta(C_{\mathbf{a}_1})$ corresponds to the Bayes risk. Consequently, Theorem 6 implies that the *DD*-classifier is equivalent to the Bayes rule under the given assumptions.

5 Implementation of *DD*-classifier

In this section we discuss several implementation issues of the *DD*-classifier.

• **Minimization of** $\hat{\Delta}_N(\mathbf{a})$

As described in Section 3, our *DD*-classifier requires searching for a polynomial with degree k_0 which minimizes the empirical misclassification rate $\hat{\Delta}_N(\mathbf{a})$ in (3.2). In principle, we need to search through all polynomials with degree k_0 which pass through the origin. However, in the linear case, i.e. $k_0 = 1$, we only need to focus on the lines which pass through the origin and at least one of the $m + n$ sample points, since all the lines running between two adjacent lines with no sample points in between would yield the same misclassification rate. Therefore, we need to consider at most $m+n$ lines. In this case, our final recommended separating line is the one that yields the minimum misclassification rate. If there are multiple such lines, we may choose the one with the smallest slope.

Similar arguments can be applied to the case of $k_0 > 1$. In other words, we now only need to consider all the polynomials which pass through the origin and k_0 of the $m + n$ sample points. Our final recommended separating polynomial is the one that yields the minimum misclassification rate.

Although the above observation simplifies significantly the search for the optimal polynomial, the computation can still be daunting when k_0 or $m + n$ are large. We notice that the difficulty involved in finding the minimum of $\hat{\Delta}_N(\mathbf{a})$ lies in the fact that the objective function $\hat{\Delta}_N(\mathbf{a})$ is the sum of many indicator functions which are not differentiable everywhere. To find a more efficient algorithm for our minimization problem, we adopt the idea in Ghosh and Chaudhuri (2005a) in using the logistic function $1/(1 + e^{-tx})$ to approximate the indicator function $I_{\{x>0\}}$ in $\hat{\Delta}_N(\mathbf{a})$. Then the minimum can be found by using appropriate derivative-based numerical methods. In this approximation, although larger t provides better approximation of $I_{\{x>0\}}$, the numerical optimization method for the resulting objective function can be rather unstable when t is large. Some care is needed in choosing t. Based on our numerical studies, we found that the optimization results become stable if we choose $t \in [50, 200]$ when the depth is standardized with upper bound 1. In all of our simulation studies and real data analysis, we chose $t = 100$.

When using numerical methods to find the minimum of the above approximation to $\hat{\Delta}_N(\mathbf{a})$, the initial value for $\mathbf{a}$ can affect the optimization procedure since the function may have many local minima. We propose the following procedure for choosing a suitable initial value. As mentioned earlier, our ideal estimate for the optimal $\mathbf{a}_0$ is the coefficient vector $\mathbf{a}$ for the polynomial which minimizes $\hat{\Delta}_N(\mathbf{a})$ among all the polynomials passing through the origin and k_0 of the sample points in the *DD*-plot. Instead of going through all these polynomials, we randomly choose sufficiently large number, say 1000, of polynomials from that set, and select the one that minimizes $\hat{\Delta}_N(\mathbf{a})$ from this subset of polynomials, and then use its coefficient vector $\mathbf{a}$ as our initial value for $\mathbf{a}$ in the numerical optimization procedure.

• **The choice of k_0**

In this paper, the degree of polynomial, k_0, is assumed to be known. However, in practice, we usually face the task of choosing the right k_0. As in the polynomial regression setting, there is a trade-off between the prediction bias and the prediction variance in the choice of k_0. The prediction here refers to the prediction of membership for the future observations based on the *DD*-classifier. Small k_0 would result in small prediction variance but large prediction bias, while large k_0 would result in small prediction bias but large prediction variance. To find a balance between these two, we recommend using cross-validation to choose k_0.

• **The choice of depth**

As shown in the simulation studies in Section 6, different depths capture different aspects of the underlying distribution. Then the *DD*-classifier can perform differently if different depths are used to construct the *DD*-plot. If some prior information about the distribution is available or if the goal is more oriented toward robustness, the suggestions given in Section 8 can offer some guidelines in choosing the appropriate depth. Otherwise, one can use a cross validation approach to choose the depth that yields the smallest misclassification rate.

6 Simulation Studies

We have conducted some simulation studies to evaluate the performance of the *DD*-classifier.

6.1 Elliptical distributions

We use simulation settings similar to those in Ghosh and Chaudhuri (2005b). Due to space limitation, we present the results for the following four settings.

(1) The underlying distributions F and G: two cases are considered. The observations are generated from bivariate normal distributions or bivariate Cauchy distributions.

(2) The dispersion matrices Σ_1 and Σ_2 for F and G: two cases are considered. If we denote $\Sigma_0 = (\sigma_{i,j})$, where $\sigma_{1,1} = \sigma_{1,2} = \sigma_{1,3} = 1$ and $\sigma_{2,2} = 4$, then, the first corresponds to the setting of equal dispersion matrices with $\Sigma_1 = \Sigma_2 = \Sigma_0$. The second corresponds to the setting of unequal dispersion matrices with $\Sigma_1 = \Sigma_0$ and $\Sigma_2 = 4\Sigma_0$.

In all four settings, the location parameters μ_1 and μ_2 for F and G are set to be $(0, 0)'$ and $(1, 1)'$, respectively, and the sample sizes, m and n, for the training sets are both set to 200.

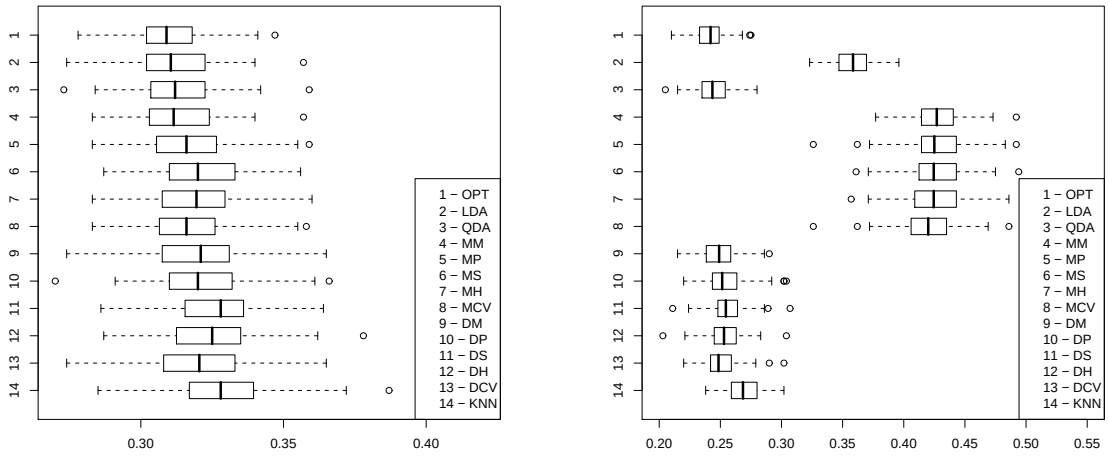
For each simulation setting, we generate a training set consisting of m and n observations from F and G, respectively. Based on this training set, various classifiers are obtained.

Another 1000 observations (500 from each group) are then generated to compute the misclassification rates for different classifiers. This experiment is repeated 100 times. The boxplots of the misclassification rates for different classifiers from the 100 experiments is then used to summarize the simulation results.

The classifiers considered in our simulation comparison studies are the linear classifier from linear discriminant analysis (denoted by LDA), quadratic classifier from quadratic discriminant analysis (denoted by QDA), k-nearest neighbor classifier (denoted by *KNN*), maximum depth classifier studied in Ghosh and Chaudhuri (2005b), and our *DD*-classifier. For *KNN* classifier, we use the leave-one-out cross validation to choose the optimal k. Since the maximum depth classifier and our *DD*-classifier both depend on the choice of depths, we consider the depths mentioned in Section 2 in our simulation studies. In the plots, the maximum depth classifiers with Mahalanobis depth, projection depth, half-space depth and simplicial depth are denoted by MM, MP, MH and MS, respectively. Similarly, the resulting *DD*-classifiers with those four depths are denoted by DM, DP, DH and DS in the plots. For each of our *DD*-classifiers paired with these four different depths, 10-fold cross validation is used to choose the optimal degree of polynomial, k_0, among 1, 2 and 3. The choice of depth in the maximum depth classifier and our *DD*-classifier can be also made by cross validations. We denote those classifiers using the depth selected from the 10-fold cross validation by MCV and DCV, respectively. We also include the optimal Bayes rule (denoted by OPT) as a benchmark for the performance comparison of different classifiers.

Figure 4 shows the boxplots of the misclassification rates of various classifiers when F and G are bivariate normal distributions. The left panel is for the case where F and G have only a location difference. In this case, all the classifiers perform similarly. All the depth-based classifiers are comparable with the optimal classifier, since both maximum depth classifier and *DD*-classifier are asymptotically equivalent to the optimal Bayes rule in this case. Among all the *DD*-classifiers, DM, as expected, is slightly better than others, and the cross validation successfully chooses DM 60 times out of 100.

Figure 4(b) shows the classification results when the dispersion matrices of F and G are also different. Here, QDA has the best performance, and is asymptotically equivalent to the optimal Bayes rule. The maximum depth classifiers perform much worse than the optimum, which can be explained similarly by Figure 2(a). In contrast, our *DD*-classifiers perform much better, all are comparable to QDA. Again, in this Gaussian case, DM performs better than other *DD*-classifiers and the cross validation chooses DM 45 times out of 100. Finally, we note that all our *DD*-classifiers outperform *KNN*.

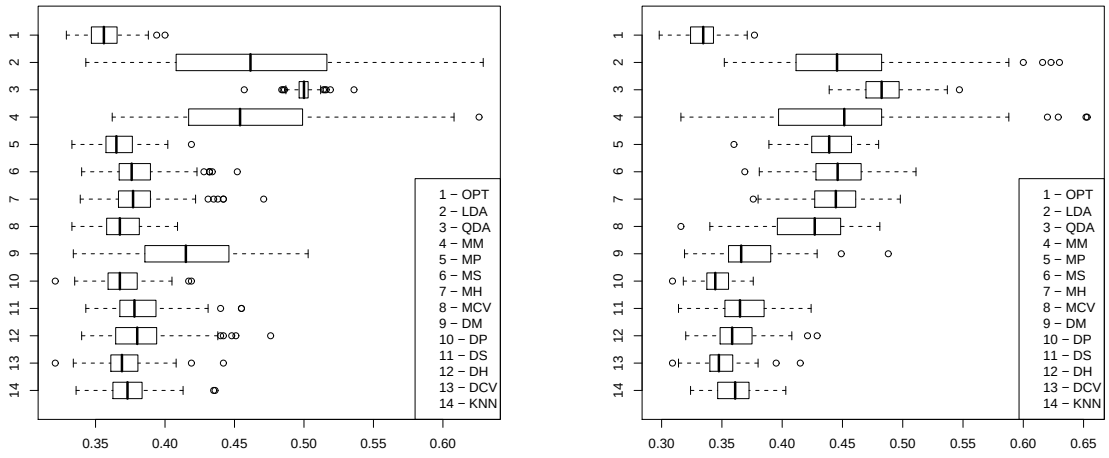


(a) (b)

Figure 4: Boxplots of misclassification rates – F and G are bivariate normal distributions with $\mu_1 = (0,0)'$ and $\mu_2 = (1,1)'$: (a) $\Sigma_1 = \Sigma_2 = \Sigma_0$; (b) $\Sigma_1 = \Sigma_0$ and $\Sigma_2 = 4\Sigma_0$.

Figure 5 shows the boxplots of the misclassification rates when F and G are bivariate Cauchy distributions. Figure 5(a) is for the case where the two dispersion matrices are the same. In this case, both LDA and QDA perform very poorly. However, the maximum depth classifiers and the *DD*-classifiers when using the depths other than Mahalanobis depth perform very well. This is not surprising since, asymptotically, they should be equivalent to the optimal Bayes rule. The poor performance of MM and DM is mainly due to the fact that the mean and covariance matrix used in Mahalanobis depth are not well defined for Cauchy distributions. The cross validation chooses this depth only 1 and 7 times, respectively, for maximum depth classifier and *DD*-classifier. However, it chooses projection depth, which is shown to be the best choice in this setting, 69 and 75 times, respectively, for maximum depth classifier and *DD*-classifier. All *DD*-classifiers and *KNN* perform comparably.

Figure 5(b) shows the results when the two dispersion matrices are different. Again, the performance of both LDA and QDA is very poor. From the plot, we can also see that, under this setting, our *DD*-classifiers clearly outperform the maximum depth classifiers. DP seems to yield the best performance among all the *DD*-classifiers. The cross validation chooses DP 79 times out of 100. Also, DP and DCV clearly perform better than *KNN*.



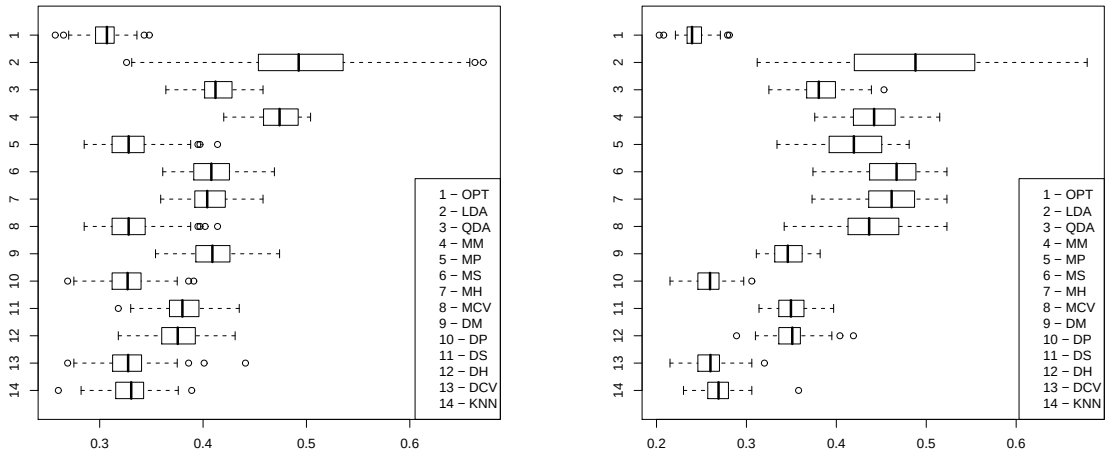
(a) (b)

Figure 5: Boxplots of misclassification rates – F and G are bivariate Cauchy distributions with $\mu_1 = (0,0)'$ and $\mu_2 = (1,1)'$: (a) $\Sigma_1 = \Sigma_2 = \Sigma_0$; (b) $\Sigma_1 = \Sigma_0$ and $\Sigma_2 = 4\Sigma_0$.

6.1.1 Robustness aspects

Here we present a simulation study to examine the robustness properties of *DD*-classifiers. The simulation settings are similar to those considered in Cui et al. (2008). Specifically, we use the earlier simulation settings in this section where F and G are both bivariate normal distributions, and 10% of the observations from F in the training set are contaminated with observations from $N(10\mu_2, \Sigma_1)$. Figure 6(a) shows the misclassification rates when F and G differ only in locations. In this case, among all the depth-based classifiers, MP and DP yield the best misclassification rates and are almost as good as the optimal Bayes rule, while MM and DM yield the worst rates. This can be explained by the fact that Mahalanobis depth is based on the sample mean and sample covariance which are not robust against outliers, while the projection depth is based on the median and MAD which are more robust statistics. Over all, *DD*-classifiers outperform their maximum depth counterparts. Both LDA and QDA perform poorly in this case. The cross validation for our *DD*-classifier chooses DP 99 times out of 100. Here, the performance of DP and DCV is similar to that of *KNN*.

With the additional dispersion difference in F and G, Figure 6(b) shows even better performance of the *DD*-classifiers than those observed in Figure 6(a). Most noticeably, MP is now much worse than DP. In this setting, DP yields the best misclassification rate among



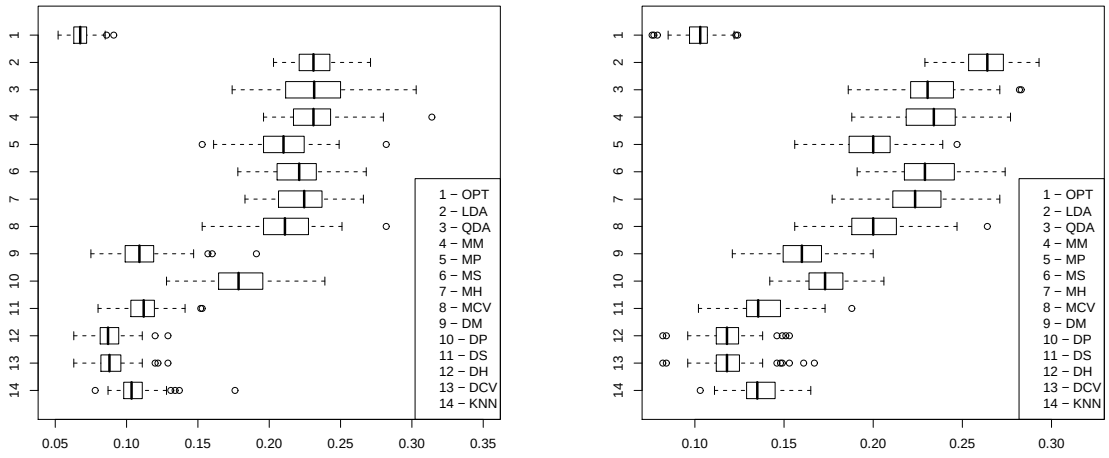
(a) (b)

Figure 6: Boxplots of misclassification rates – F and G are bivariate normal distributions with $\mu_1 = (0,0)'$ and $\mu_2 = (1,1)'$, and the training observations from F were contaminated with observations from $N(10\mu_2, \Sigma_1)$: (a) $\Sigma_1 = \Sigma_2 = \Sigma_0$; (b) $\Sigma_1 = \Sigma_0$ and $\Sigma_2 = 4\Sigma_0$.

all the classifiers. The cross validation for our *DD*-classifier chooses DP all the time. Here, the performance of DP, and thus DCV, is slightly better than that of *KNN*.

6.2 Non-elliptical distributions

We have also conducted a simulation study on some non-elliptical distributions to show the broader applicability of *DD*-classifier. To facilitate the exposition, we denote a bivariate distribution F which has independent marginal distributions by $F = (F_1, F_2)$, where F_1 and F_2 are the marginal distributions. The first two settings involve exponential distributions. Denote the exponential distribution with mean λ by $\text{Exp}(\lambda)$. In our first setting, we choose F as $(\text{Exp}(1), \text{Exp}(1))$ and G as the shifted bivariate distribution, $(\text{Exp}(1)+1, \text{Exp}(1)+1)$. Therefore, F and G differ only in the location in this setting. In the second setting, we choose F as $(\text{Exp}(1), \text{Exp}(2))$ and G as $(\text{Exp}(2)+1, \text{Exp}(1)+1)$. Figure 7 shows the classification results for these two settings.



(a) (b)

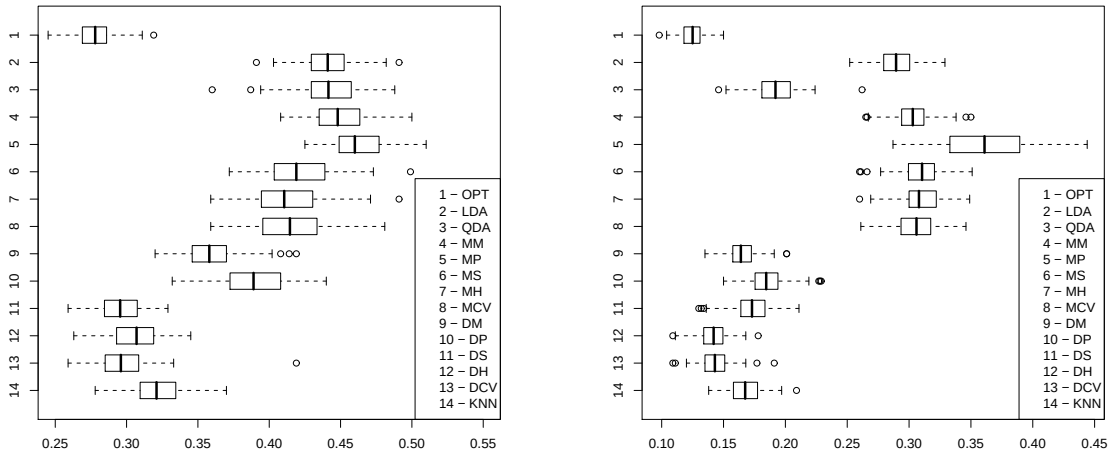
Figure 7: Boxplots of misclassification rates: (a) $F = (\text{Exp}(1), \text{Exp}(1))$ and $G = (\text{Exp}(1) + 1, \text{Exp}(1) + 1)$; (b) $F = (\text{Exp}(1), \text{Exp}(2))$ and $G = (\text{Exp}(2) + 1, \text{Exp}(1) + 1)$.

To introduce the next setting, we first denote the distribution of Z below by $\text{MixN}(\mu; \sigma_1, \sigma_2)$,

$$Z = \begin{cases} -\sigma_1 |X| + \mu & \text{with probability } 1/2 \\ \sigma_2 |X| + \mu & \text{with probability } 1/2, \end{cases}$$

where X is a standard normal random variable. In our third setting, F has a bivariate distribution $(\text{MixN}(0; 1, 2), \text{MixN}(0; 1, 4))$, and G has $(\text{MixN}(1; 1, 2), \text{MixN}(1; 1, 4))$. In our fourth setting, F is $\text{N}((0, 0)', I_2)$ and G is $(\text{Exp}(1), \text{Exp}(1))$. Figure 8 shows the results for the last two settings.

In all four settings, DS or DH outperforms all the other classifiers and they are quite comparable to the optimal Bayes rule. The cross validation here chooses the winning classifier of these two most of the time. DM and DP do not perform well in these cases. This may be explained by the fact that Mahalanobis depth and projection depth usually fail to capture non-elliptical structures of the underlying distributions. Overall, maximum depth classifiers perform poorly in all four settings. DH clearly outperforms *KNN* in all Figures 7 and 8, and DS outperforms DH in the setting of Figure 8(a). In all, DCV is consistently the best performer in all cases.



(a) (b)

Figure 8: Boxplots of misclassification rates: (a) $F = (\text{MixN}(0; 1, 2), \text{MixN}(0; 1, 4))$ and $G = (\text{MixN}(1; 1, 2), \text{MixN}(1; 1, 4))$; (b) $F = \text{N}((0, 0)', I_2)$ and $G = (\text{Exp}(1), \text{Exp}(1))$.

7 Applications to Real Data

In this section, we apply the *DD*-classifier to three real data sets. The brief descriptions of the data sets are given below.

• Biomedical data

This data set, available at http://lib.stat.cmu.edu/datasets/, was first discussed in Cox et al. (1982). It consists of 4 different blood measurements for 134 normal people and 75 carriers of a rare genetic disorder. After removing the 15 subjects which have missing values, 127 normal people and 67 carriers remain.

• Blood Transfusion data

This data set contains information of 748 blood donors randomly selected from the donor database of Blood Transfusion Service Center in Hsin-Chu City in Taiwan. It was first used by Yeh et al. (2009) and is available at http://archive.ics.uci.edu. These 748 donors were divided into two groups depending on whether or not the donor donated blood in March 2007. Out of the 748 donors, 178 did and 570 did not. After removing the two linearly correlated measurements, each donor is associated with three measurements: R (months since last donation), and F (total number of donation) and T (months since first donation).

• **Image Segmentation data**

This data set is accessible from http://archive.ics.uci.edu. It contains the pixel information of two types of images: cement or window. After removing the redundant or linearly correlated measurements, 10 measurements are used to characterize the pixel information. There are 330 cement images and 330 window images.

Since none of the data set above has well-defined training and test sets, we randomly divide each data set into two parts. One part serves as a training set and the other as a test set. More specifically, in the Biomedical data, out of the 194 subjects, we randomly choose 100 and 50 subjects from the normal and carrier groups, respectively, to form the training set and the remaining then serve as the test set. In the Blood Transfusion data, the training set consists of 400 and 100 donors randomly selected respectively from the two groups, and the remaining go to the test set. For the Image Segmentation data, 250 images randomly drawn from each group form the training set and the rest form the test set. For each data set, we carry out this random partition 100 times. The average test set misclassification rates and their standard errors for different classifiers over 100 replications are reported in Table 1. 50-fold cross validation is used to determine the degree of polynomial in our *DD*-classifiers, DM, DP and DH as well as the depth used in MCV and DCV. Leave-one-out cross validation is used to determine the optimal k in *KNN*.

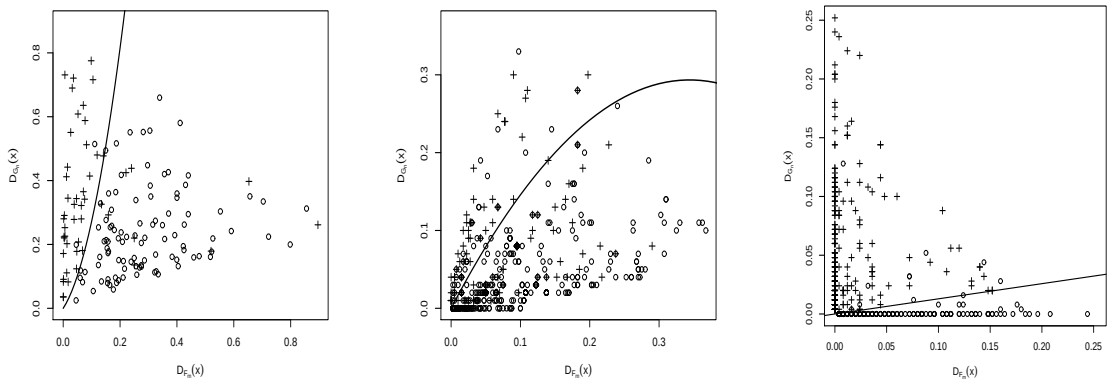
Data Sets	LDA	QDA	KNN	MM	MP	MH	MCV	DM	DP	DH	DCV
Biomedical	15.6	12.8	14.1	26.5	31.3	12.8	12.8	12.7	15.1	12.2	12.7
	(0.5)	(0.4)	(0.5)	(0.6)	(0.6)	(0.5)	(0.5)	(0.5)	(0.5)	(0.4)	(0.4)
Blood	27.7	27.7	29.0	32.5	31.5	29.6	29.6	25.6	25.8	26.3	25.7
Transfusion	(0.2)	(0.2)	(0.2)	(0.3)	(0.4)	(0.3)	(0.3)	(0.2)	(0.2)	(0.2)	(0.2)
Image	8.3	8.1	6.0	8.3	16.9	7.8	8.0	7.6	11.7	6.2	6.3
Segmentation	(0.2)	(0.2)	(0.2)	(0.2)	(0.4)	(0.2)	(0.2)	(0.2)	(0.3)	(0.2)	(0.2)

Table 1: Average misclassification rates (in percentage) with standard errors.

As shown in Table 1, in all the settings, our *DD*-classifiers clearly outperform their maximum depth counterparts. Furthermore, for Biomedical and Blood Transfusion data, the *DD*-classifiers with Mahalanobis depth and half space depth outperform LDA, QDA or *KNN*. For the Image Segmentation data, the *DD*-classifier with half-space depth holds significant edge over LDA or QDA, and is comparable to *KNN*.

The three data sets in our applications are all higher than two-dimensional, with the third being 10 dimensional. It is almost impossible to visualize directly these data sets or their

classification outcomes. On the other hand, these can be easily presented on the *DD*-plots in Figure 9. Here Figures 9(a), (b) and (c) show the *DD*-plots of the three training sets and their corresponding separating curves derived from the *DD*-classifiers. The *DD*-plots in (a), (b) and (c) are constructed using the Mahanobis depth, half-space depth and half-space depth, respectively, which are selected by a 50-fold cross validation. The curves in those plots are the *DD*-classifier polynomials, whose degrees are also selected by cross validation.



(a) (b) (c)

Figure 9: The *DD*-plots with their corresponding *DD*-classifiers: (a) Biomedical data; (b) Blood Transfusion data; (c) Image Segmentation data

Remark 8 Note that efficient exact algorithms for computing half-space depth are available only for dimensions no higher than 3, as seen in Rousseeuw and Struyf (1998). For the higher dimensional applications in this section, we use the random approximation algorithm introduced in Cuesta-Albertos and Nieto-Reyes (2008) which is computationally efficient in any dimension.

8 Concluding Remarks

In this paper, we introduce the *DD*-classifier. It is a fully nonparametric classification procedure. The classification results can always be visualized in a two-dimensional *DD*-plot no matter how large the dimension of the data is. For some settings, the proposed classifier is shown to be asymptotically equivalent to the optimal Bayes rule. In addition, our simulation studies suggest that the *DD*-classifier performs well in general settings, including

non-elliptical distributions. In many cases, the *DD*-classifier clearly outperforms the maximum depth classifier and it performs at the same or better level than other nonparametric classifiers such as the *KNN*. Perhaps equally important, our classifier is easy to implement, and it bypasses all the hassles of estimating parameters such as means, scales and so on. Overall, our *DD*-classifier is an attractive nonparametric classification approach.

Since there are different notions of data depth in the literature, our *DD*-classifier can achieve different properties by using different notions of data depth. For instance, if achieving robustness against possible sample contamination or extreme observations is the main goal, one should consider using the more robust depth such as projection depth in the *DD*-classifier. If the assumption of elliptical structure for the data is in doubt or no particular distributional information for the data is known, one should use in the *DD*-classifier a geometric depth such as simplicial depth or half-space depth to achieve better performance. This is because geometric depths generally reflect more accurately the true underlying geometric structure. In fact, sometimes the choice of depths can be made more intelligently to improve further the performance of *DD*-classifier. For example, if one sample is from a Gaussian distribution and the other sample is known to contain outliers, then Mahalanobis depth should be used for the first sample, while a robust depth such as projection depth, simplicial depth, or half-space depth should be used for the second sample.

Finally, unless pertinent information about the underlying distribution is available, we strongly recommend our proposed procedure of using cross-validation to choose the degree of the separating polynomial. This procedure clearly outperforms the linear separation when the linear separation fails to achieve the performance of the Bayes classifier, as seen in Figures 2 and 3. Even in situations where the best separating function is supposed to be linear, our proposed procedure performs just as well, as seen from Figure 4. Moreover, given the fact that the linearity of r for a given depth does not imply the linearity of r for other depths, our proposed cross-validation approach to choose the correct degree of the polynomial is more systematic and desirable. Finally, the cross-validation approach can also be a more objective approach to choosing the appropriate notion of depth in *DD*-classifiers, given that helpful indication from the data is generally unclear in practice.

9 Appendix

Proof of Proposition 1 Since D_F and D_G are strictly decreasing functions respectively of $f_1(\cdot)$ and $f_2(\cdot)$, we can write $f_1(x) = g_1\left(D_F(x)\right)$ and $f_2(x) = g_2\left(D_G(x)\right)$, where the $g_i(\cdot)$'s are some strictly increasing functions depending on the depth functions used. Therefore, the

Bayes rule is

$$\begin{cases} \pi_2 f_2(x)/\pi_1 f_1(x) > 1 & \Longleftrightarrow D_G(x) > g_2^{-1}\left[\left(\frac{\pi_1}{\pi_2}\right) g_1\left(D_F(x)\right)\right] \\ \pi_2 f_2(x)/\pi_1 f_1(x) < 1 & \Longleftrightarrow D_G(x) < g_2^{-1}\left[\left(\frac{\pi_1}{\pi_2}\right) g_1\left(D_F(x)\right)\right] \end{cases}$$

The proposition follows if we choose $r(\cdot) = g_2^{-1}\left[\left(\frac{\pi_1}{\pi_2}\right) g_1(\cdot)\right]$.

Proof of Lemma 2 Let us prove first that every sequence $\{C_n\} \subset \Gamma$ contains a subsequence $\{C_{n_k}\}$ such that there exists $C_0 \in \Gamma$ which satisfies

$$P_F\{z : C_{n_k}(D_F(z), D_G(z)) \to C_0(D_F(z), D_G(z))\} = 1 \tag{9.1}$$

$$P_G\{z : C_{n_k}(D_F(z), D_G(z)) \to C_0(D_F(z), D_G(z))\} = 1. \tag{9.2}$$

From (4.2), the result is simple if $\{C_n\}$ contains a subsequence of classification rules which are unions of at most k_0 intervals. Thus, let us assume that there exists $\mathbf{a}_n = (a_n^1, \ldots, a_n^{k_0}) \in \mathbb{R}^{k_0}$ such that $C_n = C_{\mathbf{a}_n}$ for every n. Then, $\{\mathbf{a}_n\}$ contains a subsequence (which we denote with the same notation as the initial) which satisfies one of the following statements:

a) There exists $\mathbf{a}_0 \in \mathbb{R}^{k_0}$ such that $\mathbf{a}_n \to \mathbf{a}_0$. From here, assumption (4.1) gives the result.

b) The set $A = A^+ \cup A^- := \{i : a_n^i \to +\infty\} \cup \{i : a_n^i \to -\infty\}$ is not empty and there exist $i_1, \ldots, i_h \in A$, and $M_1, \ldots, M_h \in (0, \infty)$ such that

$$\lim_n \frac{|a_n^{i_j}|}{|a_n^{i_1}|} = M_{i_j},\ j = 1, \ldots, h \text{ and } \lim_n \frac{|a_n^i|}{|a_n^{i_1}|} = 0, \text{ if } i \neq i_1, \ldots, i_h.$$

We now have three possibilities:

b.1) $A^- = \emptyset$. In this case, $r_{\mathbf{a}_n}(x) \to \infty$ for every $x \in (0, 1]$ and then $c_{\mathbf{a}_n} \to 0$.

b.2) $A^+ = \emptyset$. Thus, $r_{\mathbf{a}_n}(x) \to -\infty$ for every $x \in (0, 1]$ and then $c_{\mathbf{a}_n} \to I_{(0,1]}$.

b.3) $A^+ \neq \emptyset$ and $A^- \neq \emptyset$. Let $\mathbf{b}_n = (b_n^1, \ldots, b_n^{k_0})$, where $b_n^i = a_n^i$ if $i \in A$ and $b_n^i = 0$ if $i \notin A$. We obtain

$$\lim_n r_{\mathbf{a}_n}(x) = \lim_n |a_n^{i_1}| \frac{r_{\mathbf{b}_n}(x)}{|a_n^{i_1}|}, \text{ for every } x.$$

W.l.o.g., we can assume that $i_1 = \min(A)$. Thus, given $n \in \mathbb{N}$ and $x \in [0, 1]$, we have

$$r_{\mathbf{b}_n}(x) = |a_n^{i_1}| x^{i_1} \left(\sum_{i \in A^+} \frac{a_n^i}{|a_n^{i_1}|} x^{i - i_1} - \sum_{i \in A^-} \frac{|a_n^i|}{|a_n^{i_1}|} x^{i - i_1} \right) =: |a_n^{i_1}| x^{i_1} h_n(x).$$

Obviously

$$\lim_n h_n(x) = h(x) = \sum_{i \in A^+} M_i x^{i-i_1} - \sum_{j \in A^-} M_i x^{j-i_1},$$

and, consequently,

$$\lim_n r_{\mathbf{a}_n}(x) = \begin{cases} +\infty, & \text{if } h(x) > 0 \\ -\infty, & \text{if } h(x) < 0 \\ = 0, & \text{if } h(x) = 0. \end{cases}$$

However, h is a polynomial with degree $(\max A - i_1) \leq k_0$, and, therefore, there exist $J_1, \ldots, J_{h^*}$ disjoint intervals, with $h^* \leq k_0$, such that

$$\lim_n c_{\mathbf{a}_n}(x) = I_{\cup J_i}(x) \in \Gamma.$$

From here, (4.2) yields $C_0 = I_{\cup J_i}$ which is the classification rule we look for.

To end the proof of the lemma, let $\{C_n\} \subset \Gamma$ be a sequence such that

$$\Delta(C_n) \to \inf_{C \in \Gamma} \Delta(C). \tag{9.3}$$

Following the reasoning above, there exist a classification rule $C_0 \in \Gamma$ and a subsequence which satisfy (9.1) and (9.2). Clearly, this subsequence also satisfies (9.3). From here and the assumptions (4.1) and (4.2), taking into account that the classification rules are bounded, it is not too difficult to prove that C_0 is the desired optimum.

Proof of Lemma 4 This proof is inspired by the proof of the consistency of the k-means presented in Cuesta-Albertos and Matrán (1988). We begin with the construction of an adequate representation of the empirical distributions. This representation is based on the Skorohod representation theorem for the convergence in distribution in terms of almost sure convergent sequences.

First, we can assume that the random samples $\{X_m\}$ and $\{Y_n\}$ are defined on the probability space (Ω, σ, μ). To highlight the randomness of the quantities we have introduced previously, we also assume that this randomness depends on the chosen $\omega \in \Omega$. Thus, we can write $\{X_m(\omega)\}$ and $\{Y_n(\omega)\}$ for the sequences, and, F_m^ω and G_n^ω for the empirical distributions based on $\{X_1(\omega), \ldots, X_m(\omega)\}$ and $\{Y_1(\omega), \ldots, Y_n(\omega)\}$ respectively. We will also write $\hat{\Delta}_N^\omega$ and $\hat{C}_N^\omega$.

Now, since $\min(m, n) \to \infty$, the d-dimensional Glivenko-Cantelli theorem allows us to conclude that there exists a set $\Omega_0 \in \sigma$ such that, $\mu(\Omega_0) = 1$ and for every $\omega \in \Omega_0$ the sequences of distributions functions $\{F_m^\omega\}$ and $\{G_n^\omega\}$ converge, respectively, to F and G. If

we apply Skorohod's representation theorem to those sequences, then there exists a second probability space $(\mathcal{X}, \sigma_\chi, \mu_\chi)$, such that, if $\omega \in \Omega_0$, then there exist sequences of random variables $\{Z_m^{\omega,F}\}_{m\geq 0}$ and $\{Z_n^{\omega,G}\}_{n\geq 0}$ such that

(i) For every $m, n = 1, \ldots$, the distribution of $Z_m^{\omega,F}$ is F_m^ω, and the distribution of $Z_n^{\omega,G}$ is G_n^ω. Moreover, the distribution of $Z_0^{\omega,F}$ is F, and the distribution of $Z_0^{\omega,G}$ is G.

(ii) The sequences $\{Z_m^{\omega,F}\}_{m\geq 1}$ and $\{Z_n^{\omega,G}\}_{n\geq 1}$ converge almost surely to the random variables $Z_0^{\omega,F}$ and $Z_0^{\omega,G}$, respectively.

Let Ω_1 be the probability one set in which the convergences (4.3) are satisfied. Thus, the set $\Omega^* = \Omega_0 \cap \Omega_1$ has probability one. Using the above construction, if $\omega \in \Omega^*$, we have that

$$\begin{aligned}
\hat{\Delta}_N^\omega(C) &= \pi_1 P_{F_m^\omega}\left\{x:\ C\left(D_{F_m^\omega}(x), D_{G_n^\omega}(x)\right) = 1\right\} + \pi_2 P_{G_n^\omega}\left\{x:\ C\left(D_{F_m^\omega}(x), D_{G_n^\omega}(x)\right) = 0\right\} \\
&= \pi_1 \mu_\chi\left\{C\left(D_{F_m^\omega}(Z_m^{\omega,F}), D_{G_n^\omega}(Z_m^{\omega,F})\right) = 1\right\} + \pi_2 \mu_\chi\left\{C\left(D_{F_m^\omega}(Z_n^{\omega,G}), D_{G_n^\omega}(Z_n^{\omega,G})\right) = 0\right\} \\
&= S_m^\omega + T_n^\omega.
\end{aligned}$$

Let us focus on the term S_m^ω, since T_n^ω can be analyzed similarly. Obviously

$$\left|D_{F_m^\omega}(Z_m^{\omega,F}) - D_F(Z_0^{\omega,F})\right| = \left|D_{F_m^\omega}(Z_m^{\omega,F}) - D_F(Z_m^{\omega,F})\right| + \left|D_F(Z_m^{\omega,F}) - D_F(Z_0^{\omega,F})\right| = S_{m,1}^\omega + S_{m,2}^\omega.$$

The sequence $\left\{S_{m,1}^\omega\right\}_m$ converges to zero because the point ω being considered belongs to Ω_1. On the other hand, $Z_m^{\omega,F}$ are random variables defined on $\mathcal{X}$ and μ_χ-a.s. converge to $Z_0^{\omega,F}$ because $\omega \in \Omega_0$. Since D_F is continuos, we have that $S_{m,2}^\omega$ μ_χ-a.s. converges to 0. Employing the similar argument for $D_{G_n^\omega}(Z_m^{\omega,F})$, we would obtain that

$$\left(D_{F_m^\omega}(Z_m^{\omega,F}), D_{G_n^\omega}(Z_m^{\omega,F})\right) \overset{a.s.}{\to} \left(D_F(Z_0^{\omega,F}), D_G(Z_0^{\omega,F})\right),$$

where this a.s.-convergence is with respect to the probability μ_χ.

Notice that the boundary of the set $\{(u,v) : C(u,v) = 0\}$ is $\{(u,v) : u = r(v)\}$ in the case in which of $C = C_r$ with r being a polynomial and a set like $\{\{\delta_i\} \times [0,1] : i = 1, \ldots, h+1\}$, when C is the indicator of a union of h intervals. Thus, since the distribution of $Z_0^{\omega,F}$ is F, the assumption (4.1) (or (4.2), depending on the kind of C we have) and the Portmanteau Theorem, imply that

$$S_m^\omega \to \pi_1 \mu_\chi\left\{C\left(D_F(Z_0^{\omega,F}), D_G(Z_0^{\omega,F})\right) = 1\right\} = \pi_1 P_F\left\{x : C(D_F(x), D_G(x)) = 1\right\}. \quad (9.4)$$

The proof ends because (9.4) holds for every $\omega \in \Omega^*$ which is a probability one set.

Proof of Theorem 5 From Lemma 4, there exists a probability one set Ω^* such that,

$$\hat{\Delta}_N^\omega(C_0) \to \Delta(C_0), \text{ a.s.}, \tag{9.5}$$

for every $\omega \in \Omega^*$. Moreover, according to the proof of this lemma, Ω^* contains the set Ω_0 in which the Skorohod's construction in the proof of Lemma 4 and the convergence (4.3) hold.

Now, let $\omega \in \Omega^*$ be a fixed point and let us consider the sequence $\{\hat{C}_N^\omega\}_N$. If we can prove that every subsequence of $\{\hat{C}_N^\omega\}_N$ contains a further subsequence which converges to C_0, then the whole sequence must converge to C_0 and the proof will be complete.

To prove this, let us consider a subsequence of $\{\hat{C}_N^\omega\}_N$. As shown in the proof of Lemma 2, this subsequence contains a further a.s. convergent subsequence. From now on, we will only refer to this convergent subsequence of a subsequence of $\{\hat{C}_N^\omega\}_N$, and for simplicity, we still use $\{\hat{C}_N^\omega\}_N$ to denote it. We will denote its limit by C^ω.

Let us consider the sequence $\left\{\hat{\Delta}_N^\omega(\hat{C}_N^\omega)\right\}_N$. As in the proof of Lemma 4, we have

$$\begin{aligned}\hat{\Delta}_N^\omega(\hat{C}_N^\omega) &= \pi_1\mu_\chi\left\{\hat{C}_N^\omega\left(D_{F_m^\omega}(Z_m^{\omega,F}), D_{G_n^\omega}(Z_m^{\omega,F})\right) = 1\right\} \\ &\quad +\pi_2\mu_\chi\left\{\hat{C}_N^\omega\left(D_{F_m^\omega}(Z_n^{\omega,G}), D_{G_n^\omega}(Z_n^{\omega,G})\right) = 0\right\}. \end{aligned}\tag{9.6}$$

Similarly, we also obtain

$$\left(D_{F_m^\omega}(Z_m^{\omega,F}), D_{G_n^\omega}(Z_m^{\omega,F})\right) \overset{a.s.}{\to} \left(D_F(Z_0^{\omega,F}), D_G(Z_0^{\omega,F})\right).$$

On the other hand, the fact that $\hat{C}_N^\omega \to C^\omega$ implies that if x satisfies that $C^\omega(D_F(x), D_G(x)) = 1$ (resp. $C^\omega(D_F(x), D_G(x)) = 1$) and $(D_F(x), D_G(x))$ does not belong to the boundary of the set $\{C^\omega = 1\}$, then, from an index on, $\hat{C}_N^\omega(D_F(x), D_G(x)) = 1$. Since the probability that $Z_0^{\omega,F}$ belongs to the boundary of $\{C^\omega = 1\}$ is zero, we obtain that

$$\begin{aligned} I_{\left\{\hat{C}_N^\omega\left(D_{F_m^\omega}(Z_m^{\omega,F}), D_{G_n^\omega}(Z_m^{\omega,F})\right)=1\right\}} &\overset{a.s.}{\to} I_{\left\{\hat{C}^\omega\left(D_F(Z_0^{\omega,F}), D_G(Z_0^{\omega,F})\right)=1\right\}}, \\ I_{\left\{\hat{C}_N^\omega\left(D_{F_m^\omega}(Z_m^{\omega,F}), D_{G_n^\omega}(Z_m^{\omega,F})\right)=0\right\}} &\overset{a.s.}{\to} I_{\left\{\hat{C}^\omega\left(D_F(Z_0^{\omega,F}), D_G(Z_0^{\omega,F})\right)=0\right\}}, \end{aligned}$$

where the a.s.-convergence is with respect to the probability μ_χ. Since all the random variables involved here are positive and bounded by 1, the dominated convergence theorem and (9.6) imply that, for the point ω under consideration,

$$\hat{\Delta}_N^\omega(\hat{C}_N^\omega) \to \Delta(C^\omega).$$

Finally, the convergence (9.5) holds in this ω since $\omega \in \Omega^*$. Therefore, by definitions of C_0 and $\hat{C}_N^\omega$ we have

$$\Delta(C_0) = \lim \hat{\Delta}_N^\omega(C_0) \geq \lim \hat{\Delta}_N^\omega(\hat{C}_N^\omega) = \Delta(C^\omega) \geq \Delta(C_0).$$

Thus $\Delta(C^\omega) = \Delta(C_0)$. The fact that C_0 is unique implies that $C^\omega = C_0$, and the proof ends.

Proof of Theorem 6 Proposition 1 implies that $C_0 = C_{\mathbf{a}_1}$, and Theorem 5 implies that $\hat{C}_N \xrightarrow{a.s.} C_{\mathbf{a}_1}$, as $\min(m,n) \to \infty$. On the other hand, observe that

$$\begin{aligned}\left|\Delta_N(\hat{C}_N) - \Delta(C_{\mathbf{a}_1})\right| \leq & \pi_1 \int \left| I_{\{\hat{C}_N(D_{F_m}(z), D_{G_n}(z))=1\}} - I_{\{C_{\mathbf{a}_1}(D_F(z), D_G(z))=1\}} \right| f_1(z)\, dz \\ & + \pi_2 \int \left| I_{\{\hat{C}_N(D_{F_m}(z), D_{G_n}(z))=0\}} - I_{\{C_{\mathbf{a}_1}(D_F(z), D_G(z))=0\}} \right| f_2(z)\, dz.\end{aligned}$$

Combining with the almost sure pointwise convergence of empirical depth functions to population depth functions, it follows from the dominated convergence theorem that

$$\left|\Delta_N(\hat{C}_N) - \Delta(C_{\mathbf{a}_1})\right| \xrightarrow{a.s.} 0, \text{ as } \min(m,n) \to \infty.$$

Applying the dominated convergence theorem again, the theorem follows.

References

Arcones, M., Chen, Z. and Gine, E. (1994). Estimators related to U-processes with applications to multivariate medians: Asymptotic normality. *The Annals of Statistics* **22**, 1460-1477.

Christmann, A., Fischer, P., and Joachims, T. (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, **17**, 273–287.

Christmann, A. and Rousseeuw, P. J. (2001). Measuring overlap in binary regression. *Computational Statistics & Data Analysis*, **37**, 65–75.

Cox, L. H., Johnson, M. M., and Kafadar, K. (1982). Exposition of statistical graphics technology. *ASA Proc. Statist. Comp. Section*, 55–56.

Cuesta-Albertos, J. A. and Matrán, C. (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory & Related Fields*, **78**, 523–534.

Cuesta-Albertos, J. A. and Nieto-Reyes, A. (2008). The random Tukey depth. *Computational Statistics & Data Analysis*, **52**, 4979–4988 .

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

Donoho, D. L. (1982). *Breakdown Properties of Multivariate Location Estimators.* Ph.D. qualifying paper, Harvard University.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on half-space depth and projected outlyingness. *Annals of Statistics*, **20**, 1803–1827.

Dümbgen, L. (1992) Limit theorems for the simplicial depth. *Statistics & Probability Letters*, **14**, 119-128.

Friedman, J. H. (1996). Another approach to polychotomous classification. *Technical Report*, Department of Statistics, Stanford University.

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

Ghosh, A. K. and Chaudhuri, P. (2005b). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, **32**, 327–350.

Hodges, J. (1955). A bivariate sign test. *The Annals of Mathematical Statistics*, **26**, 523–527.

Li, J. and Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*, **19**, 686–696.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, **18**, 405–414.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, **27**, 783–840.

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

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth (with discussion). *Journal of the American Statistical Association*, **94**, 388–402.

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

Stahel, W. (1981). *Robust Schaetzungen: Infinitesmale Optimalitaet und Schaetzungen von Kovarianzmatrizen.* Ph.D. thesis, ETH Zurich.

Tukey, J. (1975). Mathematics and picturing data. *Proceedings of the 1975 International Congress of Mathematics*, **2**, 523-531.

Yeh, I. C., Yang, K. J. and Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, **36**, 5866-5871.

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

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