Bernoulli **21**(1), 2015, 62–82
DOI: 10.3150/13-BEJ561

Nonparametrically consistent depth-based classifiers

DAVY PAINDAVEINE[*] and GERMAIN VAN BEVER[**]

ECARES & Département de Mathématique, Université libre de Bruxelles, Belgium.
E-mail: [*]*dpaindav@ulb.ac.be;* [**]*gvbever@ulb.ac.be*

We introduce a class of depth-based classification procedures that are of a nearest-neighbor nature. Depth, after symmetrization, indeed provides the center-outward ordering that is necessary and sufficient to define nearest neighbors. Like all their depth-based competitors, the resulting classifiers are affine-invariant, hence in particular are insensitive to unit changes. Unlike the former, however, the latter achieve Bayes consistency under virtually any absolutely continuous distributions – a concept we call *nonparametric consistency*, to stress the difference with the stronger *universal consistency* of the standard kNN classifiers. We investigate the finite-sample performances of the proposed classifiers through simulations and show that they outperform affine-invariant nearest-neighbor classifiers obtained through an obvious standardization construction. We illustrate the practical value of our classifiers on two real data examples. Finally, we shortly discuss the possible uses of our depth-based neighbors in other inference problems.

Keywords: affine-invariance; classification procedures; nearest neighbors; statistical depth functions; symmetrization

1. Introduction

The main focus of this work is on the standard classification setup in which the observation, of the form $(\mathbf{X}, Y)$, is a random vector taking values in $\mathbb{R}^d \times \{0, 1\}$. A classifier is a function $m : \mathbb{R}^d \to \{0, 1\}$ that associates with any value $\mathbf{x}$ a predictor for the corresponding "class" Y. Denoting by $\mathbb{I}[A]$ the indicator function of the set A, the so-called Bayes classifier, defined through

$$m_{\text{Bayes}}(\mathbf{x}) = \mathbb{I}\big[\eta(\mathbf{x}) > 1/2\big], \qquad \text{with } \eta(\mathbf{x}) = P[Y = 1 \mid \mathbf{X} = \mathbf{x}], \tag{1.1}$$

is optimal in the sense that it minimizes the probability of misclassification $P[m(\mathbf{X}) \neq Y]$. Under absolute continuity assumptions, the Bayes rule rewrites

$$m_{\text{Bayes}}(\mathbf{x}) = \mathbb{I}\left[\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > \frac{\pi_0}{\pi_1}\right], \tag{1.2}$$

where $\pi_j = P[Y = j]$ and f_j denotes the pdf of $\mathbf{X}$ conditional on $[Y = j]$. Of course, empirical classifiers $\hat{m}^{(n)}$ are obtained from i.i.d. copies $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$, of $(\mathbf{X}, Y)$, and it is desirable that such classifiers are consistent, in the sense that, as $n \to \infty$, the probability of misclassification of $\hat{m}^{(n)}$, conditional on $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$, converges in probability to the probability of misclassification of the Bayes rule. If this convergence holds irrespective of the distribution of $(\mathbf{X}, Y)$, the consistency is said to be *universal*.

1350-7265 © 2015 ISI/BS

Classically, parametric approaches assume that the conditional distribution of $\mathbf{X}$ given $[Y = j]$ is multinormal with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ $(j = 0, 1)$. This gives rise to the so-called *quadratic discriminant analysis (QDA)* – or to *linear discriminant analysis (LDA)* if it is further assumed that $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$. It is standard to estimate the parameters $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ $(j = 0, 1)$ by the corresponding sample means and empirical covariance matrices, but the use of more robust estimators was recommended in many works; see, for example, Randles *et al.* [26], He and Fung [15], Dehon and Croux [4], or Hartikainen and Oja [14]. Irrespective of the estimators used, however, these classifiers fail to be consistent away from the elliptical case.

Denoting by $d_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}) = ((\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))^{1/2}$ the Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$ in the metric associated with the symmetric and positive definite matrix $\boldsymbol{\Sigma}$, it is well known that the QDA classifier rewrites

$$m_{\mathrm{QDA}}(\mathbf{x}) = \mathbb{I}\big[d_{\boldsymbol{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) < d_{\boldsymbol{\Sigma}_0}(\mathbf{x}, \boldsymbol{\mu}_0) + C\big], \tag{1.3}$$

where the constant C depends on $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$, and π_0, hence classifies $\mathbf{x}$ into Population 1 if it is sufficiently more central in Population 1 than in Population 0 (centrality, in elliptical setups, being therefore measured with respect to the geometry of the underlying equidensity contours). This suggests that *statistical depth functions*, that are mappings of the form $\mathbf{x} \mapsto D(\mathbf{x}, P)$ indicating how central $\mathbf{x}$ is with respect to a probability measure P (see Section 2.1 for a more precise definition), are appropriate tools to perform nonparametric classification. Indeed, denoting by P_j the probability measure associated with Population j $(j = 0, 1)$, (1.3) makes it natural to consider classifiers of the form

$$m_D(\mathbf{x}) = \mathbb{I}\big[D(\mathbf{x}, P_1) > D(\mathbf{x}, P_0)\big],$$

based on some fixed statistical depth function D. This *max-depth approach* was first proposed in Liu, Parelius and Singh [23] and was then investigated in Ghosh and Chaudhuri [13]. Dutta and Ghosh [10,11] considered max-depth classifiers based on the projection depth and on (an affine-invariant version of) the L^p depth, respectively. Hubert and Van der Veeken [17] modified the max-depth approach based on projection depth to better cope with possibly skewed data.

Recently, Li, Cuesta-Albertos and Liu [21] proposed the "Depth vs Depth" (DD) classifiers that extend the max-depth ones by constructing appropriate polynomial separating curves in the DD-plot, that is, in the scatter plot of the points $(D_0^{(n)}(\mathbf{X}_i), D_1^{(n)}(\mathbf{X}_i))$, $i = 1, \dots, n$, where $D_j^{(n)}(\mathbf{X}_i)$ refers to the depth of $\mathbf{X}_i$ with respect to the data points coming from Population j. Those separating curves are chosen to minimize the empirical misclassification rate on the training sample and their polynomial degree m is chosen through cross-validation. Lange, Mosler and Mozharovskyi [20] defined modified DD-classifiers that are computationally efficient and apply in higher dimensions (up to $d = 20$). Other depth-based classifiers were proposed in Jörnsten [18], Ghosh and Chaudhuri [12] and Cui, Lin and Yang [5].

Being based on depth, these classifiers are clearly of a nonparametric nature. An important requirement in nonparametric classification, however, is that consistency holds as broadly as possible and, in particular, does not require "structural" distributional assumptions. In that respect, the depth-based classifiers available in the literature are not so satisfactory, since they are at best consistent under elliptical distributions only.[1] This restricted-to-ellipticity consistency implies that, as far as consistency is concerned, the Mahalanobis depth is perfectly sufficient and is by no means inferior to the "more nonparametric" (Tukey [32]) halfspace depth or (Liu [22]) simplicial depth, despite the fact that it uninspiringly leads to LDA through the max-depth approach. Also, even this restricted consistency often requires estimating densities; see, for example, Dutta and Ghosh [10,11]. This is somewhat undesirable since density and depth are quite antinomic in spirit (a deepest point may very well be a point where the density vanishes). Actually, if densities are to be estimated in the procedure anyway, then it would be more natural to go for density estimation all the way, that is, to plug density estimators in (1.2).

The poor consistency of the available depth-based classifiers actually follows from their *global* nature. Zakai and Ritov [35] indeed proved that any universally consistent classifier needs to be of a *local* nature. In this paper, we therefore introduce local depth-based classifiers, that rely on nearest-neighbor ideas (kernel density techniques should be avoided, since, as mentioned above, depth and densities are somewhat incompatible). From their nearest-neighbor nature, they will inherit consistency under very mild conditions, while from their depth nature, they will inherit affine-invariance and robustness, two important features in multivariate statistics and in classification in particular. Identifying nearest neighbors through depth will be achieved via an original symmetrization construction. The corresponding depth-based neighborhoods are of a nonparametric nature and the good finite-sample behavior of the resulting classifiers most likely results from their data-driven adaptive nature.

The outline of the paper is as follows. In Section 2, we first recall the concept of statistical depth functions (Section 2.1) and then describe our symmetrization construction that allows to define the depth-based neighbors to be used later for classification purposes (Section 2.2). In Section 3, we define the proposed depth-based nearest-neighbor classifiers and present some of their basic properties (Section 3.1) before providing consistency results (Section 3.2). In Section 4, Monte Carlo simulations are used to compare the finite-sample performances of our classifiers with those of their competitors. In Section 5, we show the practical value of the proposed classifiers on two real-data examples. We then discuss in Section 6 some further applications of our depth-based neighborhoods. Finally, the Appendix collects the technical proofs.

2. Depth-based neighbors

In this section, we review the concept of statistical depth functions and define the depth-based neighborhoods on which the proposed nearest-neighbor classifiers will be based.

2.1. Statistical depth functions

Statistical depth functions allow to measure *centrality* of any $\mathbf{x} \in \mathbb{R}^d$ with respect to a probability measure P over $\mathbb{R}^d$ (the larger the depth of $\mathbf{x}$, the more central $\mathbf{x}$ is with respect to P). Following

[1] The classifiers from Dutta and Ghosh [11] are an exception that slightly extends consistency to (a subset of) the class of L_p-elliptical distributions.

Zuo and Serfling [37], we define a statistical depth function as a bounded mapping $D(\cdot, P)$ from $\mathbb{R}^d$ to $\mathbb{R}^+$ that satisfies the following four properties:

(P1) *affine-invariance*: for any $d \times d$ invertible matrix $\mathbf{A}$, any d-vector $\mathbf{b}$ and any distribution P over $\mathbb{R}^d$, $D(\mathbf{A}\mathbf{x} + \mathbf{b}, P^{\mathbf{A},\mathbf{b}}) = D(\mathbf{x}, P)$, where $P^{\mathbf{A},\mathbf{b}}$ is defined through $P^{\mathbf{A},\mathbf{b}}[B] = P[\mathbf{A}^{-1}(B - \mathbf{b})]$ for any d-dimensional Borel set B;

(P2) *maximality at center*: for any P that is symmetric about $\boldsymbol{\theta}$ (in the sense[2] that $P[\boldsymbol{\theta} + B] = P[\boldsymbol{\theta} - B]$ for any d-dimensional Borel set B), $D(\boldsymbol{\theta}, P) = \sup_{\mathbf{x} \in \mathbb{R}^d} D(\mathbf{x}, P)$;

(P3) *monotonicity relative to the deepest point*: for any P having deepest point $\boldsymbol{\theta}$, for any $\mathbf{x} \in \mathbb{R}^d$ and any $\lambda \in [0, 1]$, $D(\mathbf{x}, P) \leq D((1 - \lambda)\boldsymbol{\theta} + \lambda\mathbf{x}, P)$;

(P4) *vanishing at infinity*: for any P, $D(\mathbf{x}, P) \to 0$ as $\|\mathbf{x}\| \to \infty$.

For any statistical depth function and any $\alpha > 0$, the set $R_\alpha(P) = \{\mathbf{x} \in \mathbb{R}^d : D(\mathbf{x}, P) \geq \alpha\}$ is called *the depth region of order* α. These regions are nested, and, clearly, inner regions collect points with larger depth. Below, it will often be convenient to rather index these regions by their probability content: for any $\beta \in [0, 1)$, we will denote by $R^\beta(P)$ the smallest $R_\alpha(P)$ that has P-probability larger than or equal to β. Throughout, subscripts and superscripts for depth regions are used for depth levels and probability contents, respectively.

Celebrated instances of statistical depth functions include

(i) the Tukey [32] halfspace depth $D_H(\mathbf{x}, P) = \inf_{\mathbf{u} \in \mathcal{S}^{d-1}} P[\mathbf{u}'(\mathbf{X} - \mathbf{x}) \geq 0]$, where $\mathcal{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$ is the unit sphere in $\mathbb{R}^d$;

(ii) the Liu [22] simplicial depth $D_S(\mathbf{x}, P) = P[\mathbf{x} \in S(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{d+1})]$, where $S(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{d+1})$ denotes the closed simplex with vertices $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{d+1}$ and where $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{d+1}$ are i.i.d. P;

(iii) the Mahalanobis depth $D_M(\mathbf{x}, P) = 1/(1 + d^2_{\boldsymbol{\Sigma}(P)}(\mathbf{x}, \boldsymbol{\mu}(P)))$, for some affine-equivariant location and scatter functionals $\boldsymbol{\mu}(P)$ and $\boldsymbol{\Sigma}(P)$;

(iv) the projection depth $D_{Pr}(\mathbf{x}, P) = 1/(1 + \sup_{\mathbf{u} \in \mathcal{S}^{d-1}} |\mathbf{u}'\mathbf{x} - \mu(P_{[\mathbf{u}]})|/\sigma(P_{[\mathbf{u}]}))$, where $P_{[\mathbf{u}]}$ denotes the probability distribution of $\mathbf{u}'\mathbf{X}$ when $\mathbf{X} \sim P$ and where $\mu(P)$ and $\sigma(P)$ are univariate location and scale functionals, respectively.

Other depth functions are the simplicial volume depth, the spatial depth, the L_p depth, etc. Of course, not all such depths fulfill properties (P1)–(P4) for any distribution P; see Zuo and Serfling [37]. A further concept of depth, of a slightly different (L_2) nature, is the so-called *zonoid depth*; see Koshevoy and Mosler [19].

Of course, if d-variate observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are available, then sample versions of the depths above are simply obtained by replacing P with the corresponding empirical distribution $P^{(n)}$ (the sample simplicial depth then has a U-statistic structure).

A crucial fact for our purposes is that a sample depth provides a *center-outward ordering* of the observations with respect to the corresponding deepest point $\hat{\boldsymbol{\theta}}^{(n)}$: one may indeed order the $\mathbf{X}_i$'s in such a way that

$$D\big(\mathbf{X}_{(1)}, P^{(n)}\big) \geq D\big(\mathbf{X}_{(2)}, P^{(n)}\big) \geq \cdots \geq D\big(\mathbf{X}_{(n)}, P^{(n)}\big). \tag{2.1}$$

[2] Zuo and Serfling [37] also considers more general symmetry concepts; however, we restrict in the sequel to central symmetry, that will be the right concept for our purposes.

Neglecting possible ties, this states that, in the depth sense, $\mathbf{X}_{(1)}$ is the observation closest to $\hat{\boldsymbol{\theta}}^{(n)}$, $\mathbf{X}_{(2)}$ the second closest, ..., and $\mathbf{X}_{(n)}$ the one farthest away from $\hat{\boldsymbol{\theta}}^{(n)}$.

For most classical depths, there may be infinitely many deepest points, that form a convex region in $\mathbb{R}^d$. This will not be an issue in this work, since the symmetrization construction we will introduce, jointly with properties (Q2)–(Q3) below, asymptotically guarantees unicity of the deepest point. For some particular depth functions, unicity may even hold for finite samples: for instance, in the case of halfspace depth, it follows from Rousseeuw and Struyf [29] and results on the uniqueness of the symmetry center (Serfling [30]) that, under the assumption that the parent distribution admits a density, symmetrization implies almost sure unicity of the deepest point.

2.2. Depth-based neighborhoods

A statistical depth function, through (2.1), can be used to define neighbors of the deepest point $\hat{\boldsymbol{\theta}}^{(n)}$. Implementing a nearest-neighbor classifier, however, requires defining neighbors of any point $\mathbf{x} \in \mathbb{R}^d$. Property (P2) provides the key to the construction of an $\mathbf{x}$-outward ordering of the observations, hence to the definition of depth-based neighbors of $\mathbf{x}$: symmetrization with respect to $\mathbf{x}$.

More precisely, we propose to consider depth with respect to the empirical distribution $P_{\mathbf{x}}^{(n)}$ associated with the sample obtained by adding to the original observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ their reflections $2\mathbf{x} - \mathbf{X}_1, \ldots, 2\mathbf{x} - \mathbf{X}_n$ with respect to $\mathbf{x}$. Property (P2) implies that $\mathbf{x}$ is the – unique (at least asymptotically; see above) – deepest point with respect to $P_{\mathbf{x}}^{(n)}$. Consequently, this symmetrization construction, parallel to (2.1), leads to an ($\mathbf{x}$-outward) ordering of the form

$$D\big(\mathbf{X}_{\mathbf{x},(1)}, P_{\mathbf{x}}^{(n)}\big) \geq D\big(\mathbf{X}_{\mathbf{x},(2)}, P_{\mathbf{x}}^{(n)}\big) \geq \cdots \geq D\big(\mathbf{X}_{\mathbf{x},(n)}, P_{\mathbf{x}}^{(n)}\big).$$

Note that the reflected observations are only used to define the ordering but are not ordered themselves. For any $k \in \{1, \ldots, n\}$, this allows to identify – up to possible ties – the k nearest neighbors $\mathbf{X}_{\mathbf{x},(i)}$, $i = 1, \ldots, k$, of $\mathbf{x}$. In the univariate case ($d = 1$), these k neighbors coincide – irrespective of the statistical depth function D – with the k data points minimizing the usual distances $|X_i - x|$, $i = 1, \ldots, n$.

In the sequel, the corresponding *depth-based neighborhoods* – that is, the sample depth regions $R_{\mathbf{x},\alpha}^{(n)} = R_\alpha(P_{\mathbf{x}}^{(n)})$ – will play an important role. In accordance with the notation from the previous section, we will write $R_{\mathbf{x}}^{\beta(n)}$ for the smallest depth region $R_{\mathbf{x},\alpha}^{(n)}$ that contains at least a proportion β of the data points $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. For $\beta = k/n$, $R_{\mathbf{x}}^{\beta(n)}$ is therefore the smallest depth-based neighborhood that contains k of the $\mathbf{X}_i$'s; ties may imply that the number of data points in this neighborhood, $K_{\mathbf{x}}^{\beta(n)}$ say, is strictly larger than k.

Note that a distance (or pseudo-distance) $(\mathbf{x}, \mathbf{y}) \mapsto d(\mathbf{x}, \mathbf{y})$ that is symmetric in its arguments is not needed to identify nearest neighbors of $\mathbf{x}$. For that purpose, a collection of "distances" $\mathbf{y} \mapsto d_{\mathbf{x}}(\mathbf{y})$ from a fixed point is indeed sufficient (in particular, it is irrelevant that this distance satisfies or not the triangular inequality). In that sense, the (data-driven) symmetric distance associated with the Oja and Paindaveine [25] *lift-interdirections*, that was recently used to build nearest-neighbor regression estimators in Biau *et al.* [1], is unnecessarily strong. Also, only an ordering of the "distances" is needed to identify nearest neighbors. This *ordering* of distances *from a fixed point* $\mathbf{x}$ is exactly what the depth-based $\mathbf{x}$-outward ordering above is providing.

3. Depth-based *k*NN classifiers

In this section, we first define the proposed depth-based classifiers and present some of their basic properties (Section 3.1). We then state the main result of this paper, related to their consistency (Section 3.2).

3.1. Definition and basic properties

The standard k-nearest-neighbor (kNN) procedure classifies the point $\mathbf{x}$ into Population 1 iff there are more observations from Population 1 than from Population 0 in the smallest Euclidean ball centered at $\mathbf{x}$ that contains k data points. Depth-based kNN classifiers are naturally obtained by replacing these Euclidean neighborhoods with the depth-based neighborhoods introduced above, that is, the proposed kNN procedure classifies $\mathbf{x}$ into Population 1 iff there are more observations from Population 1 than from Population 0 in the smallest depth-based neighborhood of $\mathbf{x}$ that contains k observations – that is, in $R_{\mathbf{x}}^{\beta(n)}$, $\beta = k/n$. In other words, the proposed depth-based classifier is defined as

$$\hat{m}_D^{(n)}(\mathbf{x}) = \mathbb{I}\left[\sum_{i=1}^{n} \mathbb{I}[Y_i = 1] W_i^{\beta(n)}(\mathbf{x}) > \sum_{i=1}^{n} \mathbb{I}[Y_i = 0] W_i^{\beta(n)}(\mathbf{x})\right], \tag{3.1}$$

with $W_i^{\beta(n)}(\mathbf{x}) = \frac{1}{K_{\mathbf{x}}^{\beta(n)}} \mathbb{I}[\mathbf{X}_i \in R_{\mathbf{x}}^{\beta(n)}]$, where $K_{\mathbf{x}}^{\beta(n)} = \sum_{j=1}^{n} \mathbb{I}[\mathbf{X}_j \in R_{\mathbf{x}}^{\beta(n)}]$ still denotes the number of observations in the depth-based neighborhood $R_{\mathbf{x}}^{\beta(n)}$. Since

$$\hat{m}_D^{(n)}(\mathbf{x}) = \mathbb{I}\big[\hat{\eta}_D^{(n)}(\mathbf{x}) > 1/2\big], \qquad \text{with } \hat{\eta}_D^{(n)}(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}[Y_i = 1] W_i^{\beta(n)}(\mathbf{x}), \tag{3.2}$$

the proposed classifier is actually the one obtained by plugging, in (1.1), the depth-based estimator $\hat{\eta}_D^{(n)}(\mathbf{x})$ of the conditional expectation $\eta(\mathbf{x})$. This will be used in the proof of Theorem 3.1 below. Note that in the univariate case ($d = 1$), $\hat{m}_D^{(n)}$, irrespective of the statistical depth function D, reduces to the standard (Euclidean) kNN classifier.

It directly follows from property (P1) that the proposed classifier is affine-invariant, in the sense that the outcome of the classification will not be affected if $\mathbf{X}_1, \dots, \mathbf{X}_n$ and $\mathbf{x}$ are subject to a common (arbitrary) affine transformation. This clearly improves over the standard kNN procedure that, for example, is sensitive to unit changes. Of course, one natural way to define an affine-invariant kNN classifier is to apply the original kNN procedure on the standardized data points $\hat{\mathbf{\Sigma}}^{-1/2}\mathbf{X}_i$, $i = 1, \dots, n$, where $\hat{\mathbf{\Sigma}}$ is an affine-equivariant estimator of shape – in the sense that

$$\hat{\mathbf{\Sigma}}(\mathbf{A}\mathbf{X}_1 + \mathbf{b}, \dots, \mathbf{A}\mathbf{X}_n + \mathbf{b}) \propto \mathbf{A}\hat{\mathbf{\Sigma}}(\mathbf{X}_1, \dots, \mathbf{X}_n)\mathbf{A}'$$

for any invertible $d \times d$ matrix $\mathbf{A}$ and any d-vector $\mathbf{b}$. A natural choice for $\hat{\mathbf{\Sigma}}$ is the regular covariance matrix, but more robust choices, such as, for example, the shape estimators from Tyler [33], Dümbgen [9], or Hettmansperger and Randles [16] would allow to get rid of any moment assumption. Here, we stress that, unlike our *adaptive* depth-based methodology, such a transformation approach leads to neighborhoods that do not exploit the geometry of the distribution in the vicinity of the point $\mathbf{x}$ to be classified (these neighborhoods indeed all are ellipsoids with $\mathbf{x}$-independent orientation and shape); as we show through simulations below, this results into significantly worse performances.

Most depth-based classifiers available – among which those relying on the max-depth approach of Liu, Parelius and Singh [23] and Ghosh and Chaudhuri [13], as well as the more efficient ones from Li, Cuesta-Albertos and Liu [21] – suffer from the "outsider problem[3]": if the point $\mathbf{x}$ to be classified does not sit in the convex hull of any of the two populations, then most statistical depth functions will give $\mathbf{x}$ zero depth with respect to each population, so that $\mathbf{x}$ cannot be classified through depth. This is of course undesirable, all the more so that such a point $\mathbf{x}$ may very well be easy to classify. To improve on this, Hoberg and Mosler [24] proposed extending the original depth fields by using the Mahalanobis depth outside the supports of both populations, a solution that quite unnaturally requires combining two depth functions. Quite interestingly, our symmetrization construction implies that the depth-based kNN classifier (that involves one depth function only) does not suffer from the outsider problem; this is an important advantage over competing depth-based classifiers.

While our depth-based classifiers in (3.1) are perfectly well-defined and enjoy, as we will show in Section 3.2 below, excellent consistency properties, practitioners might find quite arbitrary that a point $\mathbf{x}$ such that $\sum_{i=1}^n \mathbb{I}[Y_i = 1] W_i^{\beta(n)}(\mathbf{x}) = \sum_{i=1}^n \mathbb{I}[Y_i = 0] W_i^{\beta(n)}(\mathbf{x})$ is assigned to Population 0. Parallel to the standard kNN classifier, the classification may alternatively be based on the population of the next neighbor. Since ties are likely to occur when using depth, it is natural to rather base classification on the proportion of data points from each population in the next depth region. Of course, if the next depth region still leads to an ex-aequo, the outcome of the classification is to be determined on the subsequent depth regions, until a decision is reached (in the unlikely case that an ex-aequo occurs for all depth regions to be considered, classification should then be done by flipping a coin). This treatment of ties is used whenever real or simulated data are considered below.

Finally, practitioners have to choose some value for the smoothing parameter k_n. This may be done, for example, through cross-validation (as we will do in the real data example of Section 5). The value of k_n is likely to have a strong impact on finite-sample performances, as confirmed in the simulations we conduct in Section 4.

3.2. Consistency results

As expected, the local (nearest-neighbor) nature of the proposed classifiers makes them consistent under very mild conditions. This, however, requires that the statistical depth function D satisfies the following further properties:

[3] The term "outsider" was recently introduced in Lange, Mosler and Mozharovskyi [20].

(Q1) *continuity*: if P is symmetric about $\boldsymbol{\theta}$ and admits a density that is positive at $\boldsymbol{\theta}$, then $\mathbf{x} \mapsto D(\mathbf{x}, P)$ is continuous in a neighborhood of $\boldsymbol{\theta}$;

(Q2) *unique maximization at the symmetry center*: if P is symmetric about $\boldsymbol{\theta}$ and admits a density that is positive at $\boldsymbol{\theta}$, then $D(\boldsymbol{\theta}, P) > D(\mathbf{x}, P)$ for all $\mathbf{x} \neq \boldsymbol{\theta}$;

(Q3) *consistency*: for any bounded d-dimensional Borel set B, $\sup_{\mathbf{x}\in B} |D(\mathbf{x}, P^{(n)}) - D(\mathbf{x}, P)| = \mathrm{o}(1)$ almost surely as $n \to \infty$, where $P^{(n)}$ denotes the empirical distribution associated with n random vectors that are i.i.d. P.

Property (Q2) complements property (P2), and, in view of property (P3), only further requires that $\boldsymbol{\theta}$ is a strict local maximizer of $\mathbf{x} \mapsto D(\mathbf{x}, P)$. Note that properties (Q1)–(Q2) jointly ensure that the depth-based neighborhoods of $\mathbf{x}$ from Section 2.2 collapse to the singleton $\{\mathbf{x}\}$ when the depth level increases to its maximal value. Finally, since our goal is to prove that our classifier satisfies an asymptotic property (namely, consistency), it is not surprising that we need to control the asymptotic behavior of the sample depth itself (property (Q3)). As shown by Theorem A.1, properties (Q1)–(Q3) are satisfied for many classical depth functions.

We can now state the main result of the paper, that shows that, unlike their depth-based competitors (that at best are consistent under semiparametric – typically elliptical – distributional assumptions), the proposed classifiers achieve consistency under virtually any absolutely continuous distributions. We speak of *nonparametric consistency*, in order to stress the difference with the stronger *universal consistency* of the standard kNN classifiers.

Theorem 3.1. *Let D be a depth function satisfying* (P2), (P3) *and* (Q1)–(Q3). *Let k_n be a sequence of positive integers such that $k_n \to \infty$ and $k_n = \mathrm{o}(n)$ as $n \to \infty$. Assume that, for $j = 0, 1$, $\mathbf{X} \mid [Y = j]$ admits a density f_j whose collection of discontinuity points has Lebesgue measure zero. Then the depth-based k_nNN classifier $m_D^{(n)}$ in* (3.1) *is consistent in the sense that*

$$P\big[m_D^{(n)}(\mathbf{X}) \neq Y \mid \mathcal{D}_n\big] - P\big[m_{\mathrm{Bayes}}(\mathbf{X}) \neq Y\big] = \mathrm{o}_P(1) \qquad \text{as } n \to \infty,$$

where $\mathcal{D}_n$ is the sigma-algebra associated with $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$.

Classically, consistency results for classification are based on a famous theorem from Stone [31]; see, for example, Theorem 6.3 in Devroye, Györfi and Lugosi [6]. However, it is an open question whether condition (i) of this theorem holds or not for the proposed classifiers, at least for some particular statistical depth functions. A sufficient condition for condition (i) is actually that there exists a partition of $\mathbb{R}^d$ into cones $C_1, \ldots, C_{\gamma_d}$ with vertex at the origin of $\mathbb{R}^d$ (γ_d not depending on n) such that, for any $\mathbf{X}_i$ and any j, there exist (with probability one) at most k data points $\mathbf{X}_\ell \in \mathbf{X}_i + C_j$ that have $\mathbf{X}_i$ among their k depth-based nearest neighbors. Would this be established for some statistical depth function D, it would prove that the corresponding depth-based k_nNN classifier $\hat{m}_D^{(n)}$ is *universally consistent*, in the sense that consistency holds without *any* assumption on the distribution of $(\mathbf{X}, Y)$.

Now, it is clear from the proof of Stone's theorem that this condition (i) may be dropped if one further assumes that $\mathbf{X}$ admits a uniformly continuous density. This is however a high price to pay, and that is the reason why the proof of Theorem 3.1 rather relies on an argument recently used in Biau *et al.* [1]; see the Appendix.

4. Simulations

We performed simulations in order to evaluate the finite-sample performances of the proposed depth-based kNN classifiers. We considered six setups, focusing on bivariate $\mathbf{X}_i$'s ($d = 2$) with equal a priori probabilities ($\pi_0 = \pi_1 = 1/2$), and involving the following densities f_0 and f_1:

Setup 1 (Multinormality). f_j, $j = 0, 1$, is the pdf of the bivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, where

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_1 = 4\boldsymbol{\Sigma}_0.$$

Setup 2 (Bivariate Cauchy). f_j, $j = 0, 1$, is the pdf of the bivariate Cauchy distribution with location center $\boldsymbol{\mu}_j$ and scatter matrix $\boldsymbol{\Sigma}_j$, with the same values of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ as in Setup 1.

Setup 3 (Flat covariance structure). f_j, $j = 0, 1$, is the pdf of the bivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, where

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_0 = \begin{pmatrix} 5^2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0.$$

Setup 4 (Uniform distributions on half-moons). f_0 and f_1 are the densities of

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} U \\ V \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} -0.5 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 & 0.5 \\ 0.5 & -1 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix},$$

respectively, where $U \sim \text{Unif}(-1, 1)$ and $V|[U = u] \sim \text{Unif}(1 - u^2, 2(1 - u^2))$;

Setup 5 (Uniform distributions on rings). f_0 and f_1 are the uniform distributions on the concentric rings $\{\mathbf{x} \in \mathbb{R}^2 : 1 \leq \|\mathbf{x}\| \leq 2\}$ and $\{\mathbf{x} \in \mathbb{R}^2 : 1.75 \leq \|\mathbf{x}\| \leq 2.5\}$, respectively.

Setup 6 (Bimodal populations). f_j, $j = 0, 1$, is the pdf of the multinormal mixture $\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_j^I, \boldsymbol{\Sigma}_j^I) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_j^{II}, \boldsymbol{\Sigma}_j^{II})$, where

$$\boldsymbol{\mu}_0^I = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\mu}_0^{II} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_0^I = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_0^{II} = 4\boldsymbol{\Sigma}_0^I,$$

$$\boldsymbol{\mu}_1^I = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}, \qquad \boldsymbol{\mu}_1^{II} = \begin{pmatrix} 4.5 \\ 4.5 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_1^I = \begin{pmatrix} 4 & 0 \\ 0 & 0.5 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_1^{II} = \begin{pmatrix} 0.75 & 0 \\ 0 & 5 \end{pmatrix}.$$

For each of these six setups, we generated 250 training and test samples of size $n = n_{\text{train}} = 200$ and $n_{\text{test}} = 100$, respectively, and evaluated the misclassification frequencies of the following classifiers:

1. The usual LDA and QDA classifiers (LDA/QDA);

2. The standard Euclidean kNN classifiers (kNN), with $\beta = k/n = 0.01, 0.05, 0.10$ and 0.40, and the corresponding “Mahalanobis” kNN classifiers (kNNaff) obtained by performing the Euclidean kNN classifiers on standardized data, where standardization is based on the regular covariance matrix estimate of the whole training sample;
3. The proposed depth-based kNN classifiers (D-kNN) for each combination of the k used in kNN/kNNaff and a statistical depth function (we focused on halfspace depth, simplicial depth, and Mahalanobis depth);
4. The depth vs depth (DD) classifiers from Li, Cuesta-Albertos and Liu [21], for each combination of a polynomial curve of degree m ($m = 1$, 2, or 3) and a statistical depth function (halfspace depth, simplicial depth, or Mahalanobis depth). Exact DD-classifiers (DD) as well as smoothed versions (DDsm) were actually implemented – although, for computational reasons, only the smoothed version was considered for $m = 3$. Exact classifiers search for the best separating polynomial curve $(d, r(d))$ of order m passing through the origin and m “DD-points” $(D_0^{(n)}(\mathbf{X}_i), D_1^{(n)}(\mathbf{X}_i))$ (see the Introduction) in the sense that it minimizes the misclassification error
$$\sum_{i=1}^{n} \bigl(\mathbb{I}[Y_i = 1]\mathbb{I}\bigl[d_i^{(n)} > 0\bigr] + \mathbb{I}[Y_i = 0]\mathbb{I}\bigl[-d_i^{(n)} > 0\bigr]\bigr), \tag{4.1}$$
with $d_i^{(n)} := r(D_0^{(n)}(\mathbf{X}_i)) - D_1^{(n)}(\mathbf{X}_i)$. Smoothed versions use derivative-based methods to find a polynomial minimizing (4.1), where the indicator $\mathbb{I}[d > 0]$ is replaced by the logistic function $1/(1 + \mathrm{e}^{-td})$ for a suitable t. As suggested in Li, Cuesta-Albertos and Liu [21], value $t = 100$ was chosen in these simulations. 100 randomly chosen polynomials were used as starting points for the minimization algorithm, the classifier using the resulting polynomial with minimal misclassification (note that this time-consuming scheme always results into better performances than the one adopted in Li, Cuesta-Albertos and Liu [21], where only one minimization is performed, starting from the best random polynomial considered).

Since the DD classification procedure is a refinement of the max-depth procedures of Ghosh and Chaudhuri [13] that leads to better misclassification rates (see Li, Cuesta-Albertos and Liu [21]), the original max-depth procedures were omitted in this study.

Boxplots of misclassification frequencies (in percentages) are reported in Figures 1 and 2. It is seen that in most setups, the proposed depth-based kNN classifiers compete well with the Euclidean kNN classifiers. The latter, however, should be avoided since (i) their outcome may unpleasantly depend on measurement units, and since (ii) the spherical nature of the neighborhoods used lead to performances that are severely affected by the – notoriously delicate – choice of k; see the “flat” Setup 3. This motivates restricting to affine-invariant classifiers, that (i) are totally insensitive to any unit changes and that (ii) can adapt to the flat structure of Setup 3 as they show there performances that are much more stable in k.

Now, regarding the comparisons between affine-invariant classifiers, the simulations results lead to the following conclusions: (i) the proposed affine-invariant depth-based classifiers outperform the natural affine-invariant versions of kNN classifiers. In other words, the natural way to make the standard kNN classifier affine-invariant results into a dramatic cost in terms of finite-sample performances. (ii) The proposed depth-based kNN classifiers also compete well with

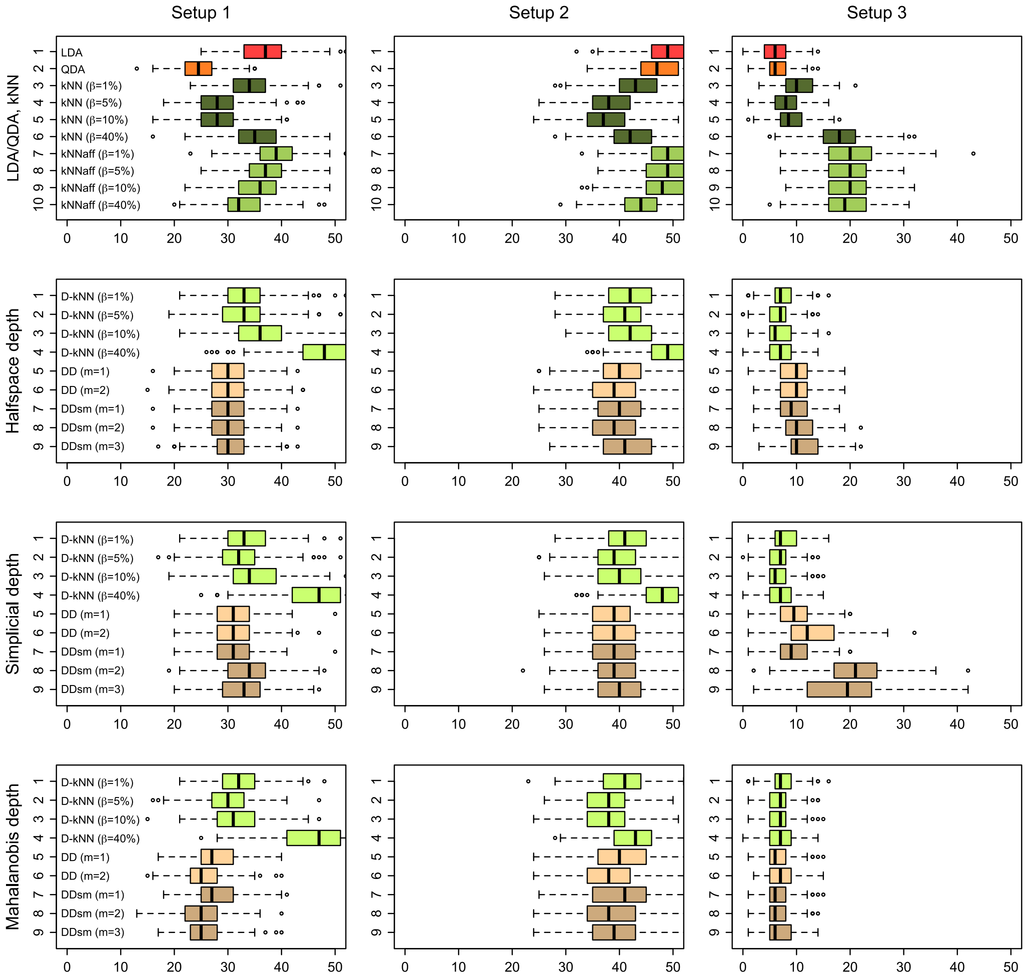

Figure 1. Boxplots of misclassification frequencies (in percentages), from 250 replications of Setups 1 to 3 described in Section 4, with training sample size $n = n_{\text{train}} = 200$ and test sample size $n_{\text{test}} = 100$, of the LDA/QDA classifiers, the Euclidean kNN classifiers (kNN) and their Mahalanobis (affine-invariant) counterparts (kNNaff), the proposed depth-based kNN classifiers (D-kNN), and some exact and smoothed version of the DD-classifiers (DD and DDsm); see Section 4 for details.

DD-classifiers both in elliptical and non-elliptical setups. Away from ellipticity (Setups 4 to 6), in particular, they perform at least as well – and sometimes outperform (Setup 4) – DD-classifiers; a single exception is associated with the use of Mahalanobis depth in Setup 5, where the DD-classifiers based on $m = 2, 3$ perform better. Apparently, another advantage of depth-based kNN classifiers over DD-classifiers is that their finite-sample performances depend much less on the statistical depth function D used.

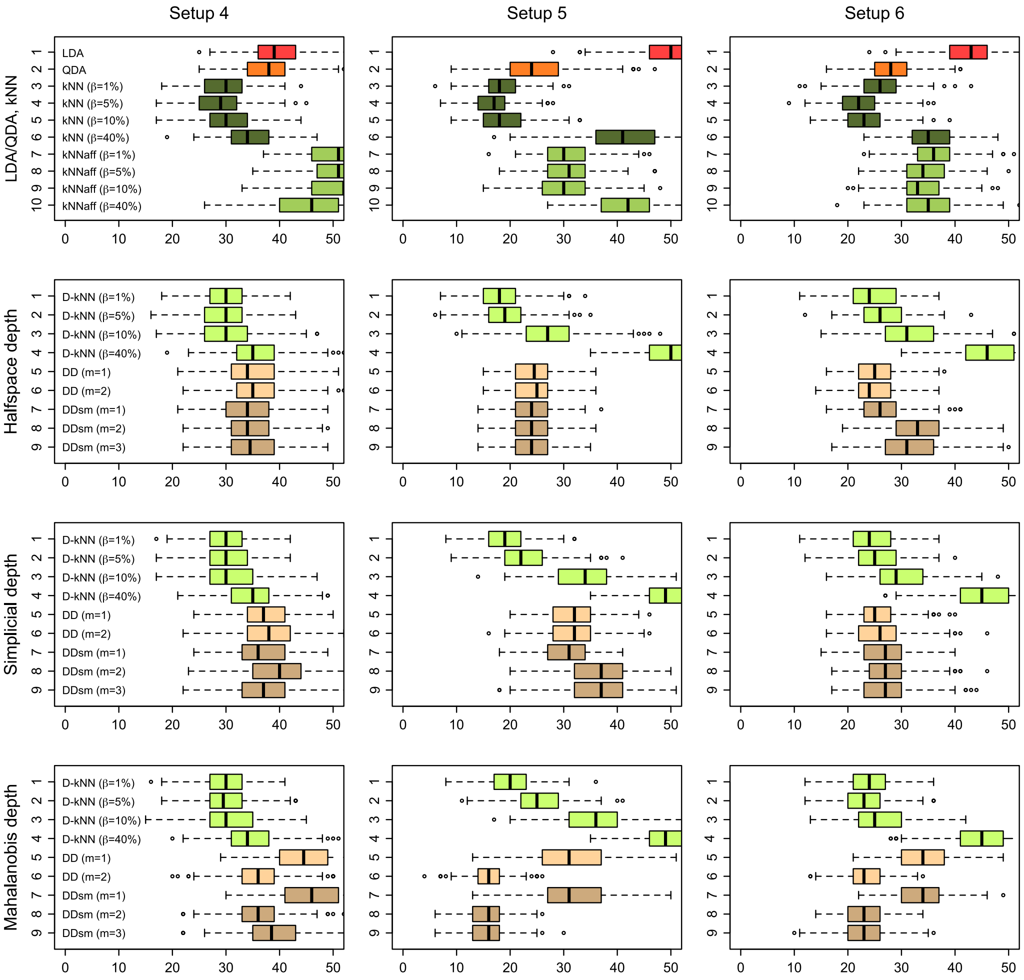

Figure 2. Boxplots of misclassification frequencies (in percentages), from 250 replications of Setups 4 to 6 described in Section 4, with training sample size $n = n_{\text{train}} = 200$ and test sample size $n_{\text{test}} = 100$, of the LDA/QDA classifiers, the Euclidean kNN classifiers (kNN) and their Mahalanobis (affine-invariant) counterparts (kNNaff), the proposed depth-based kNN classifiers (D-kNN), and some exact and smoothed version of the DD-classifiers (DD and DDsm); see Section 4 for details.

5. Real-data examples

In this section, we investigate the performances of our depth-based kNN classifiers on two well known benchmark datasets. The first example is taken from Ripley [27] and can be found on the book's website (http://www.stats.ox.ac.uk/pub/PRNN). This data set involves well-specified training and test samples, and we therefore simply report the test set misclassification rates of the different classifiers included in the study. The second example, blood transfusion data, is available at http://archive.ics.uci.edu/ml/index.html. Unlike the first data set, no clear partition into a training sample and a test sample is provided here. As suggested in Li, Cuesta-Albertos and Liu [21], we randomly performed such a partition 100 times (see the details below) and computed the average test set misclassification rates, together with standard deviations.

A brief description of each dataset is as follows:

Synthetic data was introduced and studied in Ripley [27]. The dataset is made of observations from two populations, each of them being actually a mixture of two bivariate normal distributions differing only in location. As mentioned above, a partition into a training sample and a test sample is provided: the training and test samples contain 250 and 1000 observations, respectively, and both samples are divided equally between the two populations.

Transfusion data contains the information on 748 blood donors selected from the blood donor database of the Blood Transfusion Service Center in Hsin-Chu City, Taiwan. It was studied in Yeh, Yang and Ting [34]. The classification problem at hand is to know whether or not the donor gave blood in March 2007. In this dataset, prior probabilities are not equal; out of 748 donors, 178 gave blood in March 2007, when 570 did not. Following Li, Cuesta-Albertos and Liu [21], one out of two linearly correlated variables was removed and three measurements were available for each donor: Recency (number of months since the last donation), Frequency (total number of donations) and Time (time since the first donation). The training set consists in 100 donors from the first class and 400 donors from the second, while the rest is assigned to the test sample (therefore containing 248 individuals).

Table 1 reports the – exact (synthetic) or averaged (transfusion) – misclassification rates of the following classifiers: the linear (LDA) and quadratic (QDA) discriminant rules, the standard kNN classifier (kNN) and its Mahalanobis affine-invariant version (kNNaff), the depth-based kNN classifiers using halfspace depth (D_H-kNN) and Mahalanobis depth (D_M-kNN), and the exact DD-classifiers for any combination of a polynomial order $m \in \{1, 2\}$ and a statistical depth function among the two considered for depth-based kNN classifiers, namely the halfspace depth (DD_H) and the Mahalanobis depth (DD_M) – smoothed DD-classifiers were excluded from this

Table 1. Misclassification rates (for synthetic data) and sample averages and standard deviations (in parentheses) of misclassification rates obtained from 100 random partitions of the data into training and test samples (for transfusion data)

	Synthetic	Transfusion
LDA	10.8	29.60 (0.9)
QDA	10.2	29.21 (1.5)
kNN	8.7	29.74 (2.0)
kNNaff	11.7	30.11 (2.1)
D_H-kNN	10.1	27.75 (1.6)
D_M-kNN	14.4	27.36 (1.5)
DD_H $(m = 1)$	13.4	28.26 (1.7)
DD_H $(m = 2)$	12.9	28.33 (1.6)
DD_M $(m = 1)$	17.5	31.44 (0.1)
DD_M $(m = 2)$	12.0	31.54 (0.6)

study, as their performances, which can only be worse than those of exact versions, showed much sensitivity to the smoothing parameter t; see Section 4. For all nearest-neighbor classifiers, leave-one-out cross-validation was used to determine k.

The results from Table 1 indicate that depth-based kNN classifiers perform very well in both examples. For synthetic data, the halfspace depth-based kNN classifier (10.1%) is only dominated by the standard (Euclidean) kNN procedure (8.7%). The latter, however, has to be discarded as it is dependent on scale and shape changes – in line with this, note that the "kNN classifier" applied in Dutta and Ghosh [11] is actually the kNNaff classifier (11.7%), as classification in that paper is performed on standardized data. The Mahalanobis depth-based kNN classifiers (14.4%) does not perform as well as its halfspace counterpart. For transfusion data, however, both depth-based kNN classifiers dominate their competitors.

6. Final comments

The depth-based neighborhoods we introduced are of interest in other inference problems as well. As an illustration, consider the regression problem where the conditional mean function $\mathbf{x} \mapsto m(\mathbf{x}) = E[Y \mid \mathbf{X} = \mathbf{x}]$ is to be estimated on the basis of mutually independent copies $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$ of a random vector $(\mathbf{X}, Y)$ with values in $\mathbb{R}^d \times \mathbb{R}$, or the problem of estimating the common density f of i.i.d. random d-vectors $\mathbf{X}_i$, $i = 1, \dots, n$. The classical k_nNN estimators for these problems are

$$\hat{m}^{(n)}(\mathbf{x}) = \sum_{i=1}^{n} W_i^{\beta_n(n)}(\mathbf{x}) Y_i = \frac{1}{k_n} \sum_{i=1}^{n} \mathbb{I}\big[\mathbf{X}_i \in B_{\mathbf{x}}^{\beta_n(n)}\big] Y_i, \quad \text{and} \quad \hat{f}^{(n)}(\mathbf{x}) = \frac{k_n}{n \mu_d(B_{\mathbf{x}}^{\beta_n(n)})} \tag{6.1}$$

where $\beta_n = k_n/n$, $B_{\mathbf{x}}^{\beta(n)}$ is the smallest Euclidean ball centered at $\mathbf{x}$ that contains a proportion β of the $\mathbf{X}_i$'s, and μ_d stands for the Lebesgue measure on $\mathbb{R}^d$. Our construction naturally leads to considering the depth-based k_nNN estimators $\hat{m}_D^{(n)}(\mathbf{x})$ and $\hat{f}_D^{(n)}(\mathbf{x})$ obtained by replacing in (6.1) the Euclidean neighborhoods $B_{\mathbf{x}}^{\beta_n}$ with their depth-based counterparts $R_{\mathbf{x}}^{\beta_n(n)}$ and $k_n = \sum_{i=1}^n \mathbb{I}[\mathbf{X}_i \in B_{\mathbf{x}}^{\beta_n(n)}]$ with $K_{\mathbf{x}}^{\beta_n(n)} = \sum_{i=1}^n \mathbb{I}[\mathbf{X}_i \in R_{\mathbf{x}}^{\beta_n(n)}]$.

A thorough investigation of the properties of these depth-based procedures is of course beyond the scope of the present paper. It is, however, extremely likely that the excellent consistency properties obtained in the classification problem extend to these nonparametric regression and density estimation setups. Now, recent works in density estimation indicate that using non-spherical (actually, ellipsoidal) neighborhoods may lead to better finite-sample properties; see, for example, Chacón [2] or Chacón, Duong and Wand [3]. In that respect, the depth-based kNN estimators above are very promising since they involve non-spherical (and for most classical depth, even non-ellipsoidal) neighborhoods whose shape is determined by the local geometry of the sample. Note also that depth-based neighborhoods only require choosing a single scalar bandwidth parameter (namely, k_n), whereas general d-dimensional ellipsoidal neighborhoods impose selecting $d(d+1)/2$ bandwidth parameters.

Appendix: Proofs

The main goal of this appendix is to prove Theorem 3.1. We will need the following lemmas.

Lemma A.1. *Assume that the depth function D satisfies* (P2), (P3), (Q1), *and* (Q2). *Let P be a probability measure that is symmetric about $\boldsymbol{\theta}$ and admits a density that is positive at $\boldsymbol{\theta}$. Then,* (i) *for all $a > 0$, there exists $\alpha < \alpha_* = \max_{\mathbf{x}\in\mathbb{R}^d} D(\mathbf{x}, P)$ such that $R_\alpha(P) \subset B_{\boldsymbol{\theta}}(a) := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \boldsymbol{\theta}\| \leq a\}$;* (ii) *for all $\alpha < \alpha_*$, there exists $\xi > 0$ such that $B_{\boldsymbol{\theta}}(\xi) \subset R_\alpha(P)$.*

Proof. (i) First, note that the existence of α_* follows from property (P2). Fix then $\delta > 0$ such that $\mathbf{x} \mapsto D(\mathbf{x}, P)$ is continuous over $B_{\boldsymbol{\theta}}(\delta)$; existence of δ is guaranteed by property (Q1). Continuity implies that $\mathbf{x} \mapsto D(\mathbf{x}, P)$ reaches a minimum in $B_{\boldsymbol{\theta}}(\delta)$, and property (Q2) entails that this minimal value, α_δ say, is strictly smaller than α_*. Using property (Q1) again, we obtain that, for each $\alpha \in [\alpha_\delta, \alpha_*]$,

$$r_\alpha : \mathcal{S}^{d-1} \to \mathbb{R}^+,$$
$$\mathbf{u} \mapsto \sup\big\{r \in \mathbb{R}^+ : \boldsymbol{\theta} + r\mathbf{u} \in R_\alpha(P)\big\}$$

is a continuous function that converges pointwise to $r_{\alpha_*}(\mathbf{u}) \equiv 0$ as $\alpha \to \alpha_*$. Since $\mathcal{S}^{d-1}$ is compact, this convergence is actually uniform, that is, $\sup_{\mathbf{u}\in\mathcal{S}^{d-1}} |r_\alpha(\mathbf{u})| = o(1)$ as $\alpha \to \alpha_*$. Part (i) of the result follows.

(ii) Property (Q2) implies that, for any $\alpha \in [\alpha_\delta, \alpha_*)$, the mapping r_α takes values in $\mathbb{R}_0^+$. Therefore, there exists $\mathbf{u}_0(\alpha) \in \mathcal{S}^{d-1}$ such that $r_\alpha(\mathbf{u}) \geq r_\alpha(\mathbf{u}_0(\alpha)) = \xi_\alpha > 0$. This implies that, for all $\alpha \in [\alpha_\delta, \alpha_*)$, we have $B_{\boldsymbol{\theta}}(\xi_\alpha) \subset R_\alpha(P)$, which proves the result for these values of α. Nestedness of the $R_\alpha(P)$'s, which follows from property (P3), then establishes the result for an arbitrary $\alpha < \alpha_*$. □

Lemma A.2. *Assume that the depth function D satisfies* (P2), (P3), *and* (Q1)–(Q3). *Let P be a probability measure that is symmetric about $\boldsymbol{\theta}$ and admits a density that is positive at $\boldsymbol{\theta}$. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. P and denote by $\mathbf{X}_{\boldsymbol{\theta},(i)}$ the ith depth-based nearest neighbor of $\boldsymbol{\theta}$. Let $K_{\boldsymbol{\theta}}^{\beta_n(n)}$ be the number of depth-based nearest neighbors in $R_{\boldsymbol{\theta}}^{\beta_n}(P^{(n)})$, where $\beta_n = k_n/n$ is based on a sequence k_n that is as in Theorem 3.1 and $P^{(n)}$ stands for the empirical distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Then, for any $a > 0$, there exists $n = n(a)$ such that $\sum_{i=1}^{K_{\boldsymbol{\theta}}^{\beta_n(n)}} \mathbb{I}[\|\mathbf{X}_{\boldsymbol{\theta},(i)} - \boldsymbol{\theta}\| > a] = 0$ almost surely for all $n \geq n(a)$.*

Note that, while $\mathbf{X}_{\boldsymbol{\theta},(i)}$ may not be properly defined (because of ties), the quantity $\sum_{i=1}^{K_{\boldsymbol{\theta}}^{\beta_n(n)}} \mathbb{I}[\|\mathbf{X}_{\boldsymbol{\theta},(i)} - \boldsymbol{\theta}\| > a] = 0$ always is.

Proof of Lemma A.2. Fix $a > 0$. By Lemma A.1, there exists $\alpha < \alpha_*$ such that $R_\alpha(P) \subset B_{\boldsymbol{\theta}}(a)$. Fix then $\bar{\alpha}$ and $\varepsilon > 0$ such that $\alpha < \bar{\alpha} - \varepsilon < \bar{\alpha} + \varepsilon < \alpha_*$. Theorem 4.1 in Zuo and Serfling [38] and the fact that $P_{\boldsymbol{\theta}}^{(n)} \to P_{\boldsymbol{\theta}} = P$ weakly as $n \to \infty$ (where $P_{\boldsymbol{\theta}}^{(n)}$ and $P_{\boldsymbol{\theta}}$ are the $\boldsymbol{\theta}$-symmetrized versions of $P^{(n)}$ and P, respectively) then entail that there exists an integer n_0 such that

$$R_{\bar{\alpha}+\varepsilon}(P) \subset R_{\bar{\alpha}}\big(P_{\boldsymbol{\theta}}^{(n)}\big) \subset R_{\bar{\alpha}-\varepsilon}(P) \subset R_{\alpha}(P)$$

almost surely for all $n \geq n_0$. From Lemma A.1 again, there exists $\xi > 0$ such that $B_{\boldsymbol{\theta}}(\xi) \subset R_{\bar{\alpha}+\varepsilon}(P)$. Hence, for any $n \geq n_0$, one has that

$$B_{\boldsymbol{\theta}}(\xi) \subset R_{\bar{\alpha}}\big(P_{\boldsymbol{\theta}}^{(n)}\big) \subset B_{\boldsymbol{\theta}}(a)$$

almost surely.

Putting $N_n = \sum_{i=1}^n \mathbb{I}[\mathbf{X}_i \in B_{\boldsymbol{\theta}}(\xi)]$, the SLLN yields that $N_n/n \to P[B_{\boldsymbol{\theta}}(\xi)] = P[B_{\boldsymbol{\theta}}(\xi)] > 0$ as $n \to \infty$, since $\mathbf{X} \sim P$ admits a density that, from continuity, is positive over a neighborhood of $\boldsymbol{\theta}$. Since $k_n = \mathrm{o}(n)$ as $n \to \infty$, this implies that, for all $n \geq \tilde{n}_0$ $(\geq n_0)$,

$$\sum_{i=1}^{n} \mathbb{I}\big[\mathbf{X}_i \in R_{\bar{\alpha}}\big(P_{\boldsymbol{\theta}}^{(n)}\big)\big] \geq N_n \geq k_n$$

almost surely. It follows that, for such values of n,

$$R_{\boldsymbol{\theta}}^{\beta_n}\big(P^{(n)}\big) = R^{\beta_n}\big(P_{\boldsymbol{\theta}}^{(n)}\big) \subset R_{\bar{\alpha}}\big(P_{\boldsymbol{\theta}}^{(n)}\big) \subset B_{\boldsymbol{\theta}}(a)$$

almost surely, with $\beta_n = k_n/n$. Therefore, $\max_{i=1,\ldots,K_{\boldsymbol{\theta}}^{\beta_n(n)}} \|\mathbf{X}_{\boldsymbol{\theta},(i)} - \boldsymbol{\theta}\| \leq a$ almost surely for large n, which yields the result. □

Lemma A.3. *For a "plug-in" classification rule* $\tilde{m}^{(n)}(\mathbf{x}) = \mathbb{I}[\tilde{\eta}^{(n)}(\mathbf{x}) > 1/2]$ *obtained from a regression estimator* $\tilde{\eta}^{(n)}(\mathbf{x})$ *of* $\eta(\mathbf{x}) = E[\mathbb{I}[Y = 1] \mid \mathbf{X} = \mathbf{x}]$, *one has that* $P[\tilde{m}^{(n)}(\mathbf{X}) \neq Y] - L_{\mathrm{opt}} \leq 2(E[(\tilde{\eta}^{(n)}(\mathbf{X}) - \eta(\mathbf{X}))^2])^{1/2}$, *where* $L_{\mathrm{opt}} = P[m_{\mathrm{Bayes}}(\mathbf{X}) \neq Y]$ *is the probability of misclassification of the Bayes rule.*

Proof. Corollary 6.1 in Devroye, Györfi and Lugosi [6] states that

$$P\big[\tilde{m}^{(n)}(\mathbf{X}) \neq Y \mid \mathcal{D}_n\big] - L_{\mathrm{opt}} \leq 2E\big[\big|\tilde{\eta}^{(n)}(\mathbf{X}) - \eta(\mathbf{X})\big| \mid \mathcal{D}_n\big],$$

where $\mathcal{D}_n$ stands for the sigma-algebra associated with the training sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$. Taking expectations in both sides of this inequality and applying Jensen's inequality readily yields the result. □

Proof of Theorem 3.1. From Bayes' theorem, $\mathbf{X}$ admits the density $\mathbf{x} \mapsto f(\mathbf{x}) = \pi_0 f_0(\mathbf{x}) + \pi_1 f_1(\mathbf{x})$. Letting $\mathrm{Supp}_+(f) = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) > 0\}$ and writing $C(f_j)$ for the collection of continuity points of f_j, $j = 0, 1$, put $N = \mathrm{Supp}_+(f) \cap C(f_0) \cap C(f_1)$. Since, by assumption, $\mathbb{R}^d \setminus C(f_j)$ $(j = 0, 1)$ has Lebesgue measure zero, we have that

$$\begin{aligned} P\big[\mathbf{X} \in \mathbb{R}^d \setminus N\big] &\leq P\big[\mathbf{X} \in \mathbb{R}^d \setminus \mathrm{Supp}_+(f)\big] + \sum_{j \in \{0,1\}} P\big[\mathbf{X} \in \mathbb{R}^d \setminus C(f_j)\big] \\ &= \int_{\mathbb{R}^d \setminus \mathrm{Supp}_+(f)} f(\mathbf{x})\, \mathrm{d}x = 0, \end{aligned}$$

so that $P[\mathbf{X} \in N] = 1$. Note also that $\mathbf{x} \mapsto \eta(\mathbf{x}) = \pi_1 f_1(\mathbf{x})/(\pi_0 f_0(\mathbf{x}) + \pi_1 f_1(\mathbf{x}))$ is continuous over N.

Fix $\mathbf{x} \in N$ and let $Y_{\mathbf{x},(i)} = Y_{j(\mathbf{x})}$ with $j(\mathbf{x})$ such that $\mathbf{X}_{\mathbf{x},(i)} = \mathbf{X}_{j(\mathbf{x})}$. With this notation, the estimator $\hat{\eta}_D^{(n)}(\mathbf{x})$ from Section 3.1 rewrites

$$\hat{\eta}_D^{(n)}(\mathbf{x}) = \sum_{i=1}^{n} Y_i W_i^{\beta(n)}(\mathbf{x}) = \frac{1}{K_{\mathbf{x}}^{\beta(n)}} \sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}} Y_{\mathbf{x},(i)}.$$

Proceeding as in Biau *et al.* [1], we therefore have that (writing for simplicity β instead of β_n in the rest of the proof)

$$T^{(n)}(\mathbf{x}) := E\big[\big(\hat{\eta}_D^{(n)}(\mathbf{x}) - \eta(\mathbf{x})\big)^2\big] \le 2T_1^{(n)}(\mathbf{x}) + 2T_2^{(n)}(\mathbf{x}),$$

with

$$T_1^{(n)}(\mathbf{x}) = E\left[\left|\frac{1}{K_{\mathbf{x}}^{\beta(n)}} \sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}} \big(Y_{\mathbf{x},(i)} - \eta(\mathbf{X}_{\mathbf{x},(i)})\big)\right|^2\right]$$

and

$$T_2^{(n)}(\mathbf{x}) = E\left[\left|\frac{1}{K_{\mathbf{x}}^{\beta(n)}} \sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}} \big(\eta(\mathbf{X}_{\mathbf{x},(i)}) - \eta(\mathbf{x})\big)\right|^2\right].$$

Writing $\mathcal{D}_X^{(n)}$ for the sigma-algebra generated by $\mathbf{X}_i$, $i = 1, \ldots, n$, note that, conditional on $\mathcal{D}_X^{(n)}$, the $Y_{\mathbf{x},(i)} - \eta(\mathbf{X}_{\mathbf{x},(i)})$'s, $i = 1, \ldots, n$, are zero mean mutually independent random variables. Consequently,

$$\begin{aligned}
T_1^{(n)}(\mathbf{x}) &= E\left[\frac{1}{(K_{\mathbf{x}}^{\beta(n)})^2} \sum_{i,j=1}^{K_{\mathbf{x}}^{\beta(n)}} E\big[\big(Y_{\mathbf{x},(i)} - \eta(\mathbf{X}_{\mathbf{x},(i)})\big)\big(Y_{\mathbf{x},(j)} - \eta(\mathbf{X}_{\mathbf{x},(j)})\big) \mid \mathcal{D}_X^{(n)}\big]\right] \\
&= E\left[\frac{1}{(K_{\mathbf{x}}^{\beta(n)})^2} \sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}} E\big[\big(Y_{\mathbf{x},(i)} - \eta(\mathbf{X}_{\mathbf{x},(i)})\big)^2 \mid \mathcal{D}_X^{(n)}\big]\right] \\
&\le E\left[\frac{4}{K_{\mathbf{x}}^{\beta(n)}}\right] \le \frac{4}{k_n} = \mathrm{o}(1),
\end{aligned}$$

as $n \to \infty$, where we used the fact that $K_{\mathbf{x}}^{\beta(n)} \ge k_n$ almost surely. As for $T_2^{(n)}(\mathbf{x})$, the Cauchy–Schwarz inequality yields (for an arbitrary $a > 0$)

$$T_2^{(n)}(\mathbf{x}) \le E\left[\frac{1}{K_{\mathbf{x}}^{\beta(n)}} \sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}} \big(\eta(\mathbf{X}_{\mathbf{x},(i)}) - \eta(\mathbf{x})\big)^2\right]$$

$$
\begin{aligned}
&= E\Bigg[\frac{1}{K_{\mathbf{x}}^{\beta(n)}}\sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}}\big(\eta(\mathbf{X}_{\mathbf{x},(i)})-\eta(\mathbf{x})\big)^2\mathbb{I}\big[\|\mathbf{X}_{\mathbf{x},(i)}-\mathbf{x}\|\le a\big]\Bigg]\\
&\quad + E\Bigg[\frac{1}{K_{\mathbf{x}}^{\beta(n)}}\sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}}\big(\eta(\mathbf{X}_{\mathbf{x},(i)})-\eta(\mathbf{x})\big)^2\mathbb{I}\big[\|\mathbf{X}_{\mathbf{x},(i)}-\mathbf{x}\|> a\big]\Bigg]\\
&\le \sup_{\mathbf{x}'\in B_{\mathbf{x}}(a)}\big|\eta(\mathbf{x}')-\eta(\mathbf{x})\big|^2 + 4E\Bigg[\frac{1}{K_{\mathbf{x}}^{\beta(n)}}\sum_{i=1}^{K_{\mathbf{x}}^{\beta(n)}}\mathbb{I}\big[\|\mathbf{X}_{\mathbf{x},(i)}-\mathbf{x}\|> a\big]\Bigg]\\
&=: \tilde{T}_2(\mathbf{x};a)+\bar{T}_2^{(n)}(\mathbf{x};a).
\end{aligned}
$$

Continuity of η at $\mathbf{x}$ implies that, for any $\varepsilon > 0$, one may choose $a = a(\varepsilon) > 0$ so that $\tilde{T}_2(\mathbf{x}; a(\varepsilon)) < \varepsilon$. Since Lemma A.2 readily yields that $\bar{T}_2^{(n)}(\mathbf{x}; a(\varepsilon)) = 0$ for large n, we conclude that $T_2^{(n)}(\mathbf{x})$ – hence also $T^{(n)}(\mathbf{x})$ – is o(1). The Lebesgue dominated convergence theorem then yields that $E[(\hat{\eta}_D^{(n)}(\mathbf{X}) - \eta(\mathbf{X}))^2]$ is o(1). Therefore, using the fact that $P[\hat{m}_D^{(n)}(\mathbf{X}) \neq Y \mid \mathcal{D}_n] \ge L_{\text{opt}}$ almost surely and applying Lemma A.3, we obtain

$$
\begin{aligned}
E\big[\big|P\big[\hat{m}_D^{(n)}(\mathbf{X})\neq Y\mid\mathcal{D}_n\big]-L_{\text{opt}}\big|\big] &= E\big[P\big[\hat{m}_D^{(n)}(\mathbf{X})\neq Y\mid\mathcal{D}_n\big]-L_{\text{opt}}\big]\\
&= P\big[\hat{m}_D^{(n)}(\mathbf{X})\neq Y\big]-L_{\text{opt}} \le 2\big(E\big[\big(\hat{\eta}_D^{(n)}(\mathbf{X})-\eta(\mathbf{X})\big)^2\big]\big)^{1/2}\\
&= \mathrm{o}(1),
\end{aligned}
$$

as $n \to \infty$, which establishes the result. □

Finally, we show that properties (Q1)–(Q3) hold for several classical statistical depth functions.

Theorem A.1. *Properties* (Q1)–(Q3) *hold for* (i) *the halfspace depth and* (ii) *the simplicial depth.* (iii) *If the location and scatter functionals $\boldsymbol{\mu}(P)$ and $\boldsymbol{\Sigma}(P)$ are such that* (a) *$\boldsymbol{\mu}(P) = \boldsymbol{\theta}$ as soon as the probability measure P is symmetric about $\boldsymbol{\theta}$ and such that* (b) *the empirical versions $\boldsymbol{\mu}(P^{(n)})$ and $\boldsymbol{\Sigma}(P^{(n)})$ associated with an i.i.d. sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from P are strongly consistent for $\boldsymbol{\mu}(P)$ and $\boldsymbol{\Sigma}(P)$, then properties* (Q1)–(Q3) *also hold for the Mahalanobis depth.*

Proof. (i) The continuity of D in property (Q1) actually holds under the only assumption that P admits a density with respect to the Lebesgue measure; see Proposition 4 in Rousseeuw and Ruts [28]. Property (Q2) is a consequence of Theorems 1 and 2 in Rousseeuw and Struyf [29] and the fact that the angular symmetry center is unique for absolutely continuous distributions; see Serfling [30]. For halfspace depth, property (Q3) follows from (6.2) and (6.6) in Donoho and Gasko [7].

(ii) The continuity of D in property (Q1) actually holds under the only assumption that P admits a density with respect to the Lebesgue measure; see Theorem 2 in Liu [22]. Remark C in Liu [22] shows that, for an angularly symmetric probability measure (hence also for a centrally symmetric probability measure) admitting a density, the symmetry center is the unique point maximizing simplicial depth provided that the density remains positive in a neighborhood of the symmetry center; property (Q2) trivially follows. property (Q3) for simplicial depth is stated in Corollary 1 of Dümbgen [8].

(iii) This is trivial. □

Finally, note that properties (Q1)–(Q3) also hold for projection depth under very mild assumptions on the univariate location and scale functionals used in the definition of projection depth; see Zuo [36].

Acknowledgements

Davy Paindaveine's research is supported by an A.R.C. contract from the Communauté Française de Belgique and by the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy). Germain Van Bever's research is supported through a Mandat d'Aspirant FNRS (Fonds National pour la Recherche Scientifique), Communauté Française de Belgique. The authors are grateful to an anonymous referee and the Editor for their careful reading and insightful comments that allowed to improve the original manuscript.

References

[1] Biau, G., Devroye, L., Dujmović, V. and Krzyżak, A. (2012). An affine invariant k-nearest neighbor regression estimate. *J. Multivariate Anal.* **112** 24–34. MR2957283

[2] Chacón, J.E. (2009). Data-driven choice of the smoothing parametrization for kernel density estimators. *Canad. J. Statist.* **37** 249–265. MR2531830

[3] Chacón, J.E., Duong, T. and Wand, M.P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statist. Sinica* **21** 807–840. MR2829857

[4] Croux, C. and Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. *Canad. J. Statist.* **29** 473–493. MR1872648

[5] Cui, X., Lin, L. and Yang, G. (2008). An extended projection data depth and its applications to discrimination. *Comm. Statist. Theory Methods* **37** 2276–2290. MR2526679

[6] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. New York: Springer. MR1383093

[7] Donoho, D.L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827. MR1193313

[8] Dümbgen, L. (1992). Limit theorems for the simplicial depth. *Statist. Probab. Lett.* **14** 119–128. MR1173409

[9] Dümbgen, L. (1998). On Tyler's M-functional of scatter in high dimension. *Ann. Inst. Statist. Math.* **50** 471–491. MR1664575

[10] Dutta, S. and Ghosh, A.K. (2012). On robust classification using projection depth. *Ann. Inst. Statist. Math.* **64** 657–676. MR2880873

[11] Dutta, S. and Ghosh, A.K. (2012). On classification based on L_p depth with an adaptive choice of p. Technical Report Number R5/2011, Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India.

[12] Ghosh, A.K. and Chaudhuri, P. (2005). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli* **11** 1–27. MR2121452

[13] Ghosh, A.K. and Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scand. J. Statist.* **32** 327–350. MR2188677

[14] Hartikainen, A. and Oja, H. (2006). On some parametric, nonparametric and semiparametric discrimination rules. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **72** 61–70. Providence, RI: Amer. Math. Soc. MR2343113

[15] He, X. and Fung, W.K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *J. Multivariate Anal.* **72** 151–162. MR1740638

[16] Hettmansperger, T.P. and Randles, R.H. (2002). A practical affine equivariant multivariate median. *Biometrika* **89** 851–860. MR1946515

[17] Hubert, M. and Van der Veeken, S. (2010). Robust classification for skewed data. *Adv. Data Anal. Classif.* **4** 239–254. MR2748689

[18] Jörnsten, R. (2004). Clustering and classification based on the L_1 data depth. *J. Multivariate Anal.* **90** 67–89. MR2064937

[19] Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *Ann. Statist.* **25** 1998–2017. MR1474078

[20] Lange, T., Mosler, K. and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statist. Papers* **55** 49–69.

[21] Li, J., Cuesta-Albertos, J.A. and Liu, R.Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. *J. Amer. Statist. Assoc.* **107** 737–753. MR2980081

[22] Liu, R.Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400

[23] Liu, R.Y., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27** 783–858. MR1724033

[24] Mosler, K. and Hoberg, R. (2006). Data analysis and classification with the zonoid depth. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **72** 49–59. Providence, RI: Amer. Math. Soc. MR2343112

[25] Oja, H. and Paindaveine, D. (2005). Optimal signed-rank tests based on hyperplanes. *J. Statist. Plann. Inference* **135** 300–323. MR2200471

[26] Randles, R.H., Broffitt, J.D., Ramberg, J.S. and Hogg, R.V. (1978). Generalized linear and quadratic discriminant functions using robust estimates. *J. Amer. Statist. Assoc.* **73** 564–568.

[27] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge Univ. Press. MR1438788

[28] Rousseeuw, P.J. and Ruts, I. (1999). The depth function of a population distribution. *Metrika* **49** 213–244. MR1731769

[29] Rousseeuw, P.J. and Struyf, A. (2004). Characterizing angular symmetry and regression symmetry. *J. Statist. Plann. Inference* **122** 161–173. MR2057920

[30] Serfling, R.J. (2006). Multivariate symmetry and asymmetry. *Encyclopedia Statist. Sci.* **8** 5338–5345.

[31] Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. MR0443204

[32] Tukey, J.W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (*Vancouver, B.C.*, 1974), *Vol.* 2 523–531. Canad. Math. Congress, Montreal, Que. MR0426989

[33] Tyler, D.E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734

[34] Yeh, I.C., Yang, K.J. and Ting, T.M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Syst. Appl.* **36** 5866–5871.

[35] Zakai, A. and Ritov, Y. (2009). Consistency and localizability. *J. Mach. Learn. Res.* **10** 827–856. MR2505136

[36] Zuo, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.* **31** 1460–1490. MR2012822

[37] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005

[38] Zuo, Y. and Serfling, R. (2000). Structural properties and convergence results for contours of sample statistical depth functions. *Ann. Statist.* **28** 483–499. MR1790006

Received June 2012 and revised June 2013