MULTISCALE GEOMETRIC FEATURE EXTRACTION
FOR HIGH-DIMENSIONAL AND NON-EUCLIDEAN DATA
WITH APPLICATIONS

BY GABRIEL CHANDLER AND WOLFGANG POLONIK

Pomona College and University of California, Davis

A method for extracting multiscale geometric features from a data cloud is proposed and analyzed. Based on geometric considerations, we map each pair of data points into a real-valued feature function defined on the unit interval. Further statistical analysis is then based on the collection of feature functions. The potential of the method is illustrated by different applications, including classification and anomaly detection. Connections to other concepts, such as random set theory, localized depth measures and nonlinear dimension reduction, are also explored.

1. Introduction. Extracting qualitative features from a multivariate data cloud is an important task that receives a lot of attention in the literature. A popular approach for the construction of such methods is based on the kernel-trick (e.g. see Shawe-Taylor and Cristianini, 2004) that maps the observed $n$ data points into a usually infinite dimensional feature space.

In this work, we propose a new feature extraction method that, similar to the kernel trick, also maps points to an infinite dimensional feature space. However, rather than constructing a feature function for each data point, we construct a matrix of feature functions: A feature function for each pair of data points. By construction, these feature functions encode certain information about the geometry and shape of the point cloud. This geometric aspect underlying our construction infuses intuition to the resulting methodology that enhances interpretability. Moreover, our feature functions are real-valued functions of a one-dimensional variable, even for multivariate or infinite-dimensional data. Consequently, these functions can be plotted, resulting in a visualization tool.

Our contributions can be summarized as follows: We present a methodology for extracting multiscale information about the shape of the underlying distribution based on a sample of independent and identically distributed (iid) data of size $n$ from a Hilbert space. The approach is based on the novel idea

MSC 2010 subject classifications: Primary 62G07, 62GHXX

Keywords and phrases: Kernel-trick, local depth, kernelized depth, visualization, VC-classes, curse of dimensionality, nonlinear dimension reduction, random set
of a distribution of local depth values of a given point. Supporting visualization tools and theoretical developments are provided. The provided theory in particular shows that the underlying estimation approach to a certain extent combats the curse of dimensionality, and that there is some adaptivity to sparsity. The versatility of the proposed approach is illustrated through various applications, including classification. Implicitly, some fundamental challenges in non-parametric statistics are addressed: How to choose smoothing/tuning parameters? What is the ‘right’ scale in high dimensions? How to find informative regions or directions in high dimensions?

To heuristically describe the underlying novel idea, assume that the data lie in $\mathbb{R}^d$. For a given anchor point $x \in \mathbb{R}^d$, we construct a distribution of depth values of $x$, and the corresponding quantile function will be the feature function associated with $x$. Defining an anchor point in terms of two data points results in a matrix of feature functions.

The distribution of depths is obtained by randomly selecting subsets of $\mathbb{R}^d$ containing $x$, then finding the depths of $x$ within the subsets. The notion of depth used here is specified below. The consideration of distributions of depths was motivated by the work of Ting et al. (2013), who consider random split points in the context of what they called ‘mass estimation’.

By construction, our depth distribution is such that small quantiles (small scales) provide information about the value of the density, while large quantiles (large scales) contain information about the global depth (cf. Lemma 2.2). Intermediate scales might perhaps be most important for feature extraction, particularly in high dimensions. Finding depths within subsets is a way to localize depth. Various local depth measures have been proposed (e.g. Agostinelli and Ramanazzi 2011, Paindaveine and van Bever 2012, Dutta et al. 2016, Kotik and Hublinka 2017, Agostinelli 2018, Serfling 2019), and their multiscale nature has also been discussed. See also section 6.3.

It should also be pointed out that we do not consider the quantile level as a tuning parameter, but we consider the entire quantile function. This is similar to the concepts underlying the mode tree (Minotte and Scott, 1993), the siZer (Chaudhuri and Marron, 1999), or to Betti-curves or other representations of persistence diagrams considered in topological data analysis (TDA), e.g. Bubenik and Kim (2007), and Bubenik (2015). These methods also do not choose a tuning (smoothing) parameter, but consider all possible values of the parameter. However, for both small and large values of the tuning parameter, the resulting quantities do not contain useful information as they are degenerate. This is in stark contrast to the methodology presented here, where all the values of the tuning parameter give informative quantities. Also, the methods just mentioned summarize the entire data cloud in
essentially one function (or in a few, as in the case of Betti curves), while our approach constructs functions for all (pairs of) individual data points.

In contrast to the kernel trick, our construction is not related to a reproducing kernel Hilbert space (RKHS). Instead, our construction is more directly guided by the objective to extract geometric information. The computational burden might be somewhat higher than for RKHS-based methodologies, but given the currently available computing power, it is manageable. The algorithmic complexity of our procedure is quadratic in the sample size and linear in the dimension, meaning that a high dimension is not the major computational challenge. It turns out that, to a certain extent, the same applies to the statistical behavior.

The remainder of the paper is organized as follows. The construction of population versions of our feature functions in $\mathbb{R}^d$ are described in section 2, and corresponding empirical versions are given in section 3. Section 4 discusses generalizations to Hilbert space data. Several applications of our approach are presented in section 5. Section 6 discusses relations to other concepts, while the choice of tuning parameters is discussed in section 7. Theoretical analyses of our feature functions are presented in Section 8.

This includes robustness to the curse of dimensionality, some adaptation to sparsity, and asymptotic distribution theory, both pointwise and as a process. All the proofs are relegated to the supplementary material.

2. Distribution of depths and depth quantile functions. The basic idea underlying our approach is to define a distribution of depths of a given point $x \in \mathbb{R}^d, d \geq 2$, by randomly selecting subsets of $\mathbb{R}^d$ containing $x$ and finding the depth of $x$ within this subset. In an attempt to balance complexity and computational cost, the usual challenge in high-dimensional situations, our approach uses (right spherical) cones as subsets.

In what follows, we only consider the case $d \geq 2$. Let $F$ be a distribution on $\mathbb{R}^d$ with Lebesgue density $f$. In slight abuse of notation, we also denote by $F(t)$ with $t \in \mathbb{R}^d$ the corresponding distribution function, and below we follow the same convention for distributions denoted by other symbols. For $x \in \mathbb{R}^d$ and a vector $u$ on the unit sphere $S^{d-1}$, a line $\ell = \ell_{x,u} \subset \mathbb{R}^d$ is given by $\ell_{x,u} = \{ y = y(t) = x + tu, t \in \mathbb{R} \}$. Notice that $\ell_{x,u} = \ell_{x,-u}$.

Then, for $(x, u) \in \mathbb{R}^d \times S^{d-1}$ and $s \in \mathbb{R}$, we denote by $C_{x,u}(s)$ the (right spherical) cone with fixed opening angle $\alpha \in [0, \pi)$ that contains $x$, has $\ell_{x,u}$ as its axis of symmetry, and whose tip $y(s) = x + su$ has distance $|s|$ to $x$. Note that this includes cones opening to both sides, where the orientation of the cone $C_{x,u}(s)$ depends on the sign of $s$, where, by convention, the orientation of $C_{x,u}(0)$ is chosen to be the same as for $s > 0$. For $s \neq 0$,
we also have $C_{x,u}(s) = C_{x,-u}(-s)$. The vector $x$ is serving as our anchor point. In practice, we will have to specify a rule for how to choose an anchor point, and this rule might depend on the specific application considered (see sections 3 and 7).

We denote by $\langle \cdot, \cdot \rangle$ the standard dot-product in $\mathbb{R}^d$. Recall that $\langle z - x, u \rangle$ is the signed distance between $x \in \mathbb{R}^d$ and the projection of $z \in \mathbb{R}^d$ onto the line $\ell_{x,u}$. With $X \sim F$, we define the depth of $x$ within $C_{x,u}(s)$ as

$$d_{x,u}(s) = \min \left( P\left[ \{X \in C_{x,u}(s)\} \cap \{\langle X - x, u \rangle \leq 0\} \right], \right.$$  

$$\left. P\left[ \{X \in C_{x,u}(s)\} \cap \{\langle X - x, u \rangle \geq 0\} \right] \right).$$  

(2.1)

We can interpret $d_{x,u}(s)$ as the Tukey depth of $x$ with respect to the (in general improper) one-dimensional distribution $F_{x,u,s}$ on $\ell_{x,u}$, defined by $F_{x,u,s}(t) = P(X \in C_{x,u}(s), \langle X - x, u \rangle \leq t)$. As $x \in \mathbb{R}^d$, ‘Tukey depth’ of $x$ might be associated with half-space depth, which is not what we mean. Instead $d_{x,u}(s)$ is the (one-dimensional) Tukey depth of the origin of the one-dimensional (improper) distribution with cdf $F_{x,u,s}(t)$, i.e., we have $d_{x,u}(s) = \min \{F_{x,u,s}(0), F_{x,u,s}(\infty) - F_{x,u,s}(0)\}$, where $F_{x,u,s}(\infty) = \sup_{t \in \mathbb{R}} F_{x,u,s}(t)$ denotes the total mass of $F_{x,u,s}$. A different way of writing this is

$$d_{x,u}(s) = \min \left( F(A_{x,u}(s)), F(B_{x,u}(s)) \right),$$

where

$$A_{x,u}(s) = C_{x,u}(s) \cap H^+_{x,u} \quad \text{and} \quad B_{x,u}(s) = C_{x,u}(s) \cap H^-_{x,u},$$

(2.2)

and $H^\pm_{x,u}$ are the two closed half-spaces defined by the hyperplane $H_{x,u} = \{z \in \mathbb{R}^d : \langle z, u \rangle = \langle x, u \rangle \}$ with normal direction $u$, passing through $x$, and $H^-_{x,u}$ contains the tip $y(s) = x + su$. In other words, the hyperplane $H_{x,u}$ splits the cone $C_{x,u}(s)$ into two closed subsets $A_{x,u}(s)$ and $B_{x,u}(s)$ with $A_{x,u}(s) \cap B_{x,u}(s) = H_{x,u} \cap C_{x,u}(s)$. By definition, $A_{x,u}(s)$ contains the tip $y(s)$, i.e. $A_{x,u}(s)$ is itself a cone with base $H_{x,u} \cap C_{x,u}(s)$, and $B_{x,u}(s) = C_{x,u}(s) \setminus A_{x,u}(s)$ is a frustum. Figure 1 illustrates this representation for a given line and a given anchor point.

Notice that even for $|s| \to \infty$, we are not obtaining Tukey half-space depth of $x$ because, for a fixed $u$, we only consider parallel half-planes orthogonal to $u$. If we, however, consider all directions for a given $x$, then we can obtain (Tukey) half-space depth as $\inf_{u \in S^d} \max_{s \in \mathbb{R}} d_{x,u}(s)$. Note that $d_{x,-u}(s) = d_{x,u}(-s)$, and thus these functions depend on the direction of $u$, not just on the line $\ell_{x,u}$. This should not cause any confusion, however.

The depth quantile functions defined next are the crucial objects. Empirical versions of them will be used as feature functions (see below). Assume
that \( s \) is randomly chosen according to a distribution on \( \mathbb{R} \), meaning that we randomly choose the tip of the cone (or its distance to the anchor point \( x \) along with the orientation of the cone) on a given axis of symmetry. Formally, given a distribution \( G \) on \( \mathbb{R} \), let \( S \sim G \), and define \( q_{x,u}(\delta) \) as the quantile function of the resulting depth distribution of \( x \), i.e.

\[
q_{x,u}(\delta) = \inf \{ t \geq 0 : G(s : d_{x,u}(s) \leq t) \geq \delta \}.
\] (2.4)

The distribution \( G \) can be considered to be a tuning parameter. Often, we will assume \( G \) to be a continuous distribution; see section 7 for some discussion on the choice of \( G \).

The following two lemmas describe some properties of the depth function \( d_{x,u}(s) \) and the corresponding depth quantile function \( q_{x,u}(\delta) \). We denote by \( I_{x,u}(t) \) the sublevel sets of \( d_{x,u}(s) \), i.e.

\[
I_{x,u}(t) = \{ s \in \mathbb{R} : d_{x,u}(s) \leq t \}.
\]

**Lemma 2.1** Let \( (x, u) \in \mathbb{R}^d \times S^{d-1} \) be fixed.

(a) The function \( s \to d_{x,u}(s), s \in \mathbb{R} \) is non-increasing for \( s \leq 0 \) and non-decreasing for \( s \geq 0 \). Thus, its sublevel sets \( I_{x,u}(t), t \geq 0 \), are either empty or intervals containing 0.

(b) Let \( F \) be continuous with positive Lebesgue density on \( \mathbb{R}^d \), and let \( m_{x,u} = \sup_s d_{x,u}(s) \). The function \( s \to d_{x,u}(s), s \in \mathbb{R} \) is continuous with minimum value 0, and it is strictly decreasing on \( \{ s \leq 0 : d_{x,u}(s) < m_{x,u} \} \), and strictly increasing on \( \{ s \geq 0 : d_{x,u}(s) < m_{x,u} \} \). The intervals \( I_{x,u}(t), t \geq 0 \), are thus all closed.
(c) (rigid transformations and scale changes) Let \( \theta = (\mu, \sigma, O) \) with \( \mu \in \mathbb{R}^d, \sigma > 0 \) and \( O \in \mathbb{R}^{d \times d} \) orthonormal. If \( F_\theta(t) = F \left( \frac{t - \mu}{\sigma} \right), t \in \mathbb{R}^d \), and \( d^\theta_{x,u}(s) \) denotes the depth function corresponding to \( F_\theta \), while \( q_{x,u}(\delta) \) is the one for \( F \), then

\[
d^\theta_{x,u}(s) = d_{\sigma Ox + \mu, Ou}(\sigma s), \quad s \in \mathbb{R}.
\]

Let \( F_{x,u}(t) \) denote the cdf of the one-dimensional distribution obtained by projecting all the mass onto the line \( \ell_{x,u} \).

**Lemma 2.2** In parts (a)-(c) assume that both \( F \) and \( G \) have positive Lebesgue densities \( f \) and \( g \), respectively. Let \((x, u) \in \mathbb{R}^d \times S^{d-1} \) be fixed.

(a) With \( I_{x,u}(q_{x,u}(\delta)) = [s^l_{x,u}(\delta), s^r_{x,u}(\delta)], \delta \in (0,1), \) we have

\[
q_{x,u}(\delta) = d_{x,u}(s^l_{x,u}(\delta)) = d_{x,u}(s^r_{x,u}(\delta)).
\]

Moreover, \( q_{x,u}(\delta) \) is strictly increasing with \( q_{x,u}(0) = 0 \), and if \( g \) is symmetric about zero then \( q_{x,u}(0) = q_{x,-u}(0) \) for all \( \delta \in [0,1] \).

(b) (large scales) \( q_{x,u}(1) = \min \{ F_{x,u}(x), 1 - F_{x,u}(x) \} \) is the global Tukey depth of \( x \) for the distribution with cdf \( F_{x,u}(t) \).

(c) (small scales) \( \lim_{\delta \to 0} \frac{q_{x,u}(\delta)}{\delta^d} = c_d \frac{f(x)}{g'(|x|)} \) with \( c_d > 0 \) a known constant.

(d) (invariance) Let \( \theta = (\mu, \sigma, O) \) (as in Lemma 2.1(c)), and let \( G_\theta(t) = G(t, \sigma) \). If \( q_{x,u}(\delta) \) denotes the depth quantile function defined with \( G_\theta \) and \( F_\theta \), while \( q_{x,u}(\delta) \) is the one for \( F \), then, for any \( \theta \),

\[
q^\theta_{\sigma Ox + \mu, Ou}(\delta) = q_{x,u}(\delta).
\]

Assertion (c) makes precise the above statement that, for small values of \( \delta \), the depth quantile function \( q_{x,u}(\delta) \) contains information about the density at \( x \). Note that this localization is achieved even though we are not using local neighborhoods of \( x \). Instead, localization follows from the use of Tukey depth and the choice of cones to define the distribution of depths. (It is the angle of the cone \( 0 < \alpha < \pi \) that is important here; cf. proof of assertion (c).) Property (d) indicates a certain type of scale-invariance, provided the change in scale in \( F \) is reflected in \( G \) also. (See discussion at the end of section 3 for more on this.)

**Interpretation of depth quantile functions.** By definition, \( d_{x,u}(s) = F(C^*(s)) \) with \( C^*(s) \in \{ A_{x,u}(s), B_{x,u}(s) \} \). Thus, if we were to consider only negative or only positive values of \( s \), the function \( d_{x,u}(s) \) can be interpreted as a generalized distribution function (only evaluated on a line) of a sub-probability measure. (Notice, though, that even for fixed \((x, u)\) the sets \( C^*(s) \) might
The anchor point and the direction define the line direction given by
as follows. For each data pair $ij$ to pick both
where
\[
F_{ij}(s) = \min(F_n(A_{ij}(s)), F_n(B_{ij}(s)))
\]

Choosing the tip randomly and considering the quantile functions allows
both the combining of the negative and positive values of $s$, as well as bringing
the various ‘distribution functions’ onto the same scale. This, in turn,
makes averaging more reasonable, for instance (see section 5.2). Also note
that the quantile functions are not quantiles of the ‘distribution functions’
$d_{x,u}(s)$, but they are quantiles of the distribution of the random variable
$d_{x,u}(S)$ where $S \sim G$. As a result, the depth quantile function $q_{x,u}(\delta)$ is a
reparametrization of $d_{x,u}(s)$ (see (2.6)). A more geometric construction of
$q_{x,u}(\delta)$ is as follows. Consider the level sets of $d_{x,u}(s)$ at level $t$, which are the
intervals $I_{x,u}(t)$. Fix $\delta > 0$, and imagine increasing $t$ until the probability
content of $I_{x,u}(t)$ under $G$ equals $\delta$. Then this value $t = t_\delta$, i.e. the value of
the depth function at the limits of the interval $I_{x,u}(t_\delta)$, equals $q_{x,u}(\delta)$. (This
is (2.6).) As for comparing two quantile functions: $q_{x_1,u_1}(\delta) > q_{x_2,u_2}(\delta)$ indicates a higher mass concentration about $x_1$ than about $x_2$ (at least when ‘looking’ in directions $u_1$ and $u_2$, respectively), in the sense that there exist intervals about the respective anchor points with the same $G$-mass $\delta$ (the level sets $I_{x_1,u_1}(t_\delta)$ and $I_{x_2,u_2}(t_\delta)$, respectively), such that the sets $C^*$ corresponding to the endpoints of $I_{x_1,u_1}(t_\delta)$ carry a higher mass content than the ones corresponding to $I_{x_2,u_2}(t_\delta)$. See also the discussion on average depth
quantile functions in section 6.1.

3. Construction of the feature functions $\hat{q}_{ij}(\delta)$ for data in $\mathbb{R}^d$. Given iid data $X_1, \ldots, X_n$ from $F$, an empirical counterpart of $q_{x,u}(\delta)$ is obtained by simply replacing the distribution $F$ by the empirical distribution $F_n$ resulting in

\[
\hat{q}_{x,u}(\delta) = \inf\{t \in \mathbb{R} : G(s : \hat{d}_{x,u}(s) \leq t) \geq \delta\},
\]

where $\hat{d}_{x,u}(s) = \min(F_n(A_{x,u}(s)), F_n(B_{x,u}(s)))$. The question now is: how
to pick both $x$ and $u$. We propose to choose them depending on the data
as follows. For each data pair $X_i, X_j$, $i \neq j$, let $u_{ij} = \frac{X_i - X_j}{\|X_i - X_j\|}$ be the
direction given by $X_i$ and $X_j$, and let $m_{ij} = \frac{X_i + X_j}{2}$ be the anchor point.
The anchor point and the direction define the line $\ell_{ij} = \{y \in \mathbb{R}^d : y = y(t) = m_{ij} + tu_{ij}, t \in \mathbb{R}\}$ passing through $X_i$ and $X_j$. This gives

\[
\hat{d}_{ij}(s) = \min\left(F_n(A_{ij}(s)), F_n(B_{ij}(s))\right),
\]

\[
\hat{q}_{ij}(\delta) = \inf\{t \in \mathbb{R} : G(s : \hat{d}_{ij}(s) \leq t) \geq \delta\},
\]

The total mass of this measure equals the ‘Tukey depth of the base point $x$ among the projections of all the data on the line.’

This distribution changes with the direction $u$ and with the base point $x$. Not necessarily be nested.) The total mass of this measure equals the ‘Tukey depth of the base point $x$ among the projections of all the data on the line.’

This distribution changes with the direction $u$ and with the base point $x$. Not necessarily be nested.) The total mass of this measure equals the ‘Tukey depth of the base point $x$ among the projections of all the data on the line.’

This distribution changes with the direction $u$ and with the base point $x$. Not necessarily be nested.) The total mass of this measure equals the ‘Tukey depth of the base point $x$ among the projections of all the data on the line.’

This distribution changes with the direction $u$ and with the base point $x$. Not necessarily be nested.) The total mass of this measure equals the ‘Tukey depth of the base point $x$ among the projections of all the data on the line.’
where $A_{ij}(s)$ and $B_{ij}(s)$ are the sets $A_{x,u}(s), B_{x,u}(s)$ with $x = m_{ij}$ and $u = w_{ij}$. The corresponding empirical depth quantile function $\tilde{q}_{ij}(\delta)$ is

$$
\tilde{q}_{ij}(\delta) = \inf \{ t : G(\{ s : \hat{d}_{ij}(s) \leq t \}) \geq \delta \}.
$$

(3.2)

The functions $\tilde{q}_{ij}(\delta)$ are our feature functions. An alternative way of thinking about constructing $\hat{d}_{ij}(s)$ (see Fig. 2) is as follows. Project all the data inside

![Fig 2: Illustrates the construction of $\hat{d}_{ij}(s)$ for $d = 2$ for a 2-class classification problem (different characters correspond to different classes). Panels show one positive and one negative value of $s$. Dashed line is $\ell_{ij}$; anchor point $m_{ij}$ is red; cones $C_{ij}(s)$ are blue; projection of the data inside the cone onto $\ell_{ij}$ are indicated.](image)

$C_{ij}(s) = C_{m_{ij},w_{ij}}(s)$ onto $\ell_{ij}$, and let $\hat{d}_{ij}(s)$ be the ‘(one-dimensional) Tukey depth of $m_{ij}$ among these projections’. Similar to the population version, this means the following: Each projection of a point $X_k \in C_{ij}(s)$ can be written as $y(t_k)$, where $t_k = \langle X_k - m_{ij}, u_{ij} \rangle \in \mathbb{R}$ is the projection score, so that, in particular, $m_{ij} = y(0)$. The value $\hat{d}_{ij}(s)$ then is the (scaled by the total sample size $n$) Tukey depth of 0 among the $t_k$’s. In addition to the distribution $G$ on $\ell_{ij}$ of the tip, the angle $\alpha$ and the rule to choose the anchor point $m_{ij} \in \ell_{ij}$ are also tuning parameters. Our choice of the anchor point as the midpoint of $X_i, X_j$ was originally motivated by the application of the proposed methodology to classification (see section 5.2), though consideration of anchor points other than the actual observations appears to have some benefits, at least in high dimensions. See section 7 for some discussion of the choice of these tuning parameters.

Observe that, in analogy to the population version discussed in section 2, the empirical depth functions $\hat{d}_{ij}(s)$ and $\hat{d}_{ji}(s)$ are mirrored versions of each other. If $G$ is symmetric about zero, then they lead to the same feature functions, i.e. in this case $\tilde{q}_{ij}(\delta) = \tilde{q}_{ji}(\delta)$.

4. Depth quantile functions for object data, the kernel-trick, $Z_1-Z_2$-plots, and non-linear multi-dimensional scaling.
4.1. Object data and the kernel trick. The approach discussed above can also be applied to object data in a Hilbert space. The simple, but key, underlying observation here is that determining our depth quantile functions only requires the ability to define cones, to calculate which data points fall into the cones, and to find distances between points. All of this, of course, hinges on the existence of an inner-product. So, if our underlying data are objects \(O_1, \ldots, O_n \in \mathcal{O}\), lying in a Hilbert space \(H\) equipped with an inner-product \(\langle \cdot, \cdot \rangle_H\), then we can simply apply our methodology outlined above to the object data with this inner-product.

Interestingly, applying our methodology to functional data in \(L^2\) will map functional data into a different set of functional data, and it remains to be investigated how such an approach compares to working with the functional data directly.

Another possible application of our methodology is in combination with the kernel trick. The kernel trick consists of mapping data \(O_i\) into an RKHS \(H\) via a feature map \(\Psi\), say, such that the dot-product \(\langle \cdot, \cdot \rangle_H\) satisfies \(\langle \Psi(O_i), \Psi(O_j) \rangle_H = K(O_i, O_j)\) for some kernel \(K: \mathcal{O} \times \mathcal{O} \to \mathbb{R}\). After applying this map, we can simply apply our methodology outlined above to the transformed data \(\Psi(O_i), i = 1, \ldots, n\) with the corresponding dot-product in the RKHS. Kernel methods are available for a variety of objects, including trees, graphs, matrices, strings, tensors, functions, persistence diagrams, etc. (e.g. see Genton 2001, Shawe-Taylor and Cristianini, 2004, Cuturi 2010). Since the elements in the feature space \(\Psi(O_i)\) lie in a Hilbert space, we can apply our methodology outlined above. This can be used to investigate and to compare the geometry of the data in feature spaces.

4.2. Relation to non-linear multidimensional scaling: \(Z_1\)-\(Z_2\)-plots. Given a pair of data, say, \(O_i, O_j\) with \(\ell = \ell_{ij}\) passing through \(O_i\) and \(O_j\), and the anchor point \(m_{ij} = \frac{O_i + O_j}{2}\), all we need to determine the depth functions \(\hat{d}_{ij}(s)\), or the quantile functions \(\hat{q}_{ij}(\delta)\), is the set of two-dimensional points \((Z_{1k}^{ij}, Z_{2k}^{ij}), k = 1, \ldots, n\), where

\[
Z_{1k}^{ij} = \left\langle O_k - m_{ij}, \frac{O_i - O_j}{\|O_i - O_j\|} \right\rangle \quad \text{and} \quad Z_{2k}^{ij} = \sqrt{\|O_k - m_{ij}\|^2 - |Z_{1k}^{ij}|^2},
\]

where \(Z_{1k}^{ij}\) is the (signed) distance between \(m_{ij}\) and the projection of \(O_k\) onto \(\ell_{ij}\), and \(Z_{2k}^{ij}\) is the distance of \(O_k\) to \(\ell_{ij}\). Checking whether \(O_k\) lies in \(C_{ij}(s)\) with \(s < 0\), say, simply means checking whether \(Z_{1k}^{ij} \geq s\), and if yes, determining whether \(Z_{2k}^{ij} \geq \tan(\alpha)(Z_{1k}^{ij} - s)\). In other words, the pairs \((Z_{1k}^{ij}, Z_{2k}^{ij}) \in \mathbb{R}^2, k = 1, \ldots, n\), fully determine \(\hat{d}_{ij}(s)\), and thus \(\hat{q}_{ij}(\delta)\).
A plot of the pairs \((Z_{ij}^1, Z_{ij}^2)\), \(k \neq i, j\), is what we call a \(Z_1-Z_2\)-plot. It is some kind of non-linear projection onto two dimensions, which is in the spirit of multidimensional scaling. However, in contrast to multidimensional scaling, we do not just have one plot, but a total of \(n(n-1)/2\) such plots. Each (unordered) pair \((O_i, O_j)\) determines one \(Z_1-Z_2\)-plot, which provides a visual impression of the geometry of the data cloud if viewed in direction \(\ell_{ij}\).

One can think of the function \(\hat{q}_{ij}(\delta)\) as a summary of the \(Z_1-Z_2\)-plot. Another reason for introducing these plots is as follows. Suppose we want to perform binary classification of high-dimensional data. The collection of \(Z_1-Z_2\)-plots can be used to visually inspect whether good linear binary classification is possible. To this end let \(\ell_{ij}\) be a line connecting two points \((O_i, O_j)\) and consider the corresponding \(Z_1-Z_2\)-coordinates. Observe that a hyperplane orthogonal to the line \(\ell_{ij}\) corresponds to a vertical line in the \(Z_1-Z_2\)-plot. Thus, if we can find a pair of points \(O_i, O_j\) such that the corresponding two-dimensional \(Z_1-Z_2\)-coordinates allow good classification with a vertical line, then there exits a good binary linear classifier in the original space.

As an illustration of a possible use of the \(Z_1-Z_2\)-plots, we consider an infinite dimensional situation. Here the underlying space is the RKHS generated by the popular Gaussian radial basis function (RBF) \(K_\sigma(u, v) = \exp(-\|u-v\|^2/\sigma^2)\). The objects \(O_j\) are the corresponding feature functions created by applying the kernel-trick to two classes (Virginica vs. Versicolor) of Fisher’s famous Iris data, and we visualize the geometry of the cloud of feature functions by using the RKHS geometry as described above.

We are interested in exploring the effect on the geometry of the feature functions induced by the choice of the tuning parameter \(\sigma\). Fixing a pair \((O_i, O_j)\) and varying \(\sigma\) results in a class of \(Z_1-Z_2\) plots \(\{(Z_{1k}^i(\sigma), Z_{2k}^j(\sigma)), k = 1, \ldots, n\}\), \(\sigma > 0\). Once can also think of them as a collection of curves \(\sigma \rightarrow (Z_{1k}^i(\sigma), Z_{2k}^j(\sigma)), k = 1, \ldots, n\).

![Fig 3](image)

Fig 3: Overlaid \(Z_1-Z_2\) plots for two classes of Iris data (color coded) in feature space with Gaussian (RBF) kernel for varying \(\sigma\) based on three different pairs \((O_i, O_j)\) - left two panels \(O_i, O_j\) both in same class; right panel different classes.

Figure 3 shows such a collection of curves by varying \(\sigma\) from .3 to 3. Each ‘curve’ (extending from top to bottom) corresponds to an individual
observation as its $Z_1$-$Z_2$-coordinates change as $\sigma$ increases. Shown are within group (left and center panels) and between groups (right panel) comparisons. The first and third plot are looking in a direction which is roughly in the direction of the line connecting the two modes. The middle plot is looking roughly orthogonal to this line. The plot provides a visualization of the changing geometry in the RKHS depending on the tuning parameter $\sigma$. A visual inspection of these plots might support the choice a useful range of the tuning parameter in the kernel. A more systematic procedure to exploit these plots is desirable and still needs to be developed. The two colors (green, red) correspond to the two classes. The peculiar shape of the curves can partly be explained as follows. When $\sigma \approx 0$ (top part of the plot), all points in the feature space are essentially orthogonal as $K(u,v) \approx 0$ for $u \neq v$, and all points in the $Z_1$-$Z_2$-plot are approximately $(0, \sqrt{3}/2)$. The latter can be seen by observing that the norm of all the feature functions equals 1 (they all live on the unit sphere in the RKHS), and projecting orthogonal unit vectors onto the line $l_{ij}$ connecting two of them gives $m_{ij}$, the midpoint between the two points determining the line. The distance to this midpoint is $\sqrt{3}/2$. Since the projections equal the midpoint, their distance obviously equals zero. This then results in the point $(0, \sqrt{3}/2)$ in the $Z_1$-$Z_2$-coordinates. As $\sigma \to \infty$, the points converge to $(0,0)$ for all the distances in the RKHS converge to zero.

5. Applications. Here we illustrate possible types of applications of the depth quantile functions $\widehat{q}_{ij}(\delta)$. The goal of this exposition is to establish that these objects contain valuable information regarding the location of individual observations within a point cloud as well as the associated geometry of the point clouds. How best to use this information likely depends on the particular task at hand, and constitutes future work beyond the scope of the present paper.

For all of the tasks discussed below, we summarize the $\binom{n}{2}$ depth quantile functions by considering averages of the form $\widehat{q}_i(\delta) = n^{-1} \sum_{j \neq i} \widehat{q}_{ij}(\delta)$, or with the sum restricted to certain subsets of the data (see section 5.2). As a result, we have a new representation of our data that can be used to perform a variety of statistical tasks, such as exploratory data analysis, classification, and anomaly detection. Other statistical tasks may require combining these functions in a different way, or adapting the choice of the tuning parameters (cf section 7), which the flexibility of this idea allows. In all of the following, we use as our base distribution $G$ a uniform over a support large enough to
ensure that at the boundary of the support, all the data lie inside the cone for all \( u \) and all \( x \) under consideration. We use \( \alpha = \pi/2 \) unless otherwise noted.

5.1. **Topological data analysis, or discovering holes.** Consider observing a point cloud in \( \mathbb{R}^8 \) where the data is either generated uniformly in a ball of fixed radius or is generated uniformly in this ball with a smaller middle ball removed. We call these two supports the 8-ball and 8-annulus respectively. Specifically, our larger ball has radius 1.5 and the inner ball has radius 1.25, accounting for 25 percent of the total volume. We wonder whether, given a sample of size of \( n = 100 \) from each model, it is possible to detect if a hole is present. An application of ISOMAP (see Tenenbaum et al. 2000) appears unable to recognize the difference between the two supports. ISOMAP constructs a weighted k-nearest neighbor graph first, finds the pairwise shortest distances in this graph, and then applies MDS to these distances. Figure 4a shows one such instance with \( K = 5 \). Other values of \( K \) also were not successful. However, the depth quantile functions show quite different behavior between the two situations, and considering all intermediate values provides more information than just considering, for instance, the right hand boundary. In fact, as shown in figure 4c, the intermediate value \( \delta = .39 \) gives perfect separation of the two classes, unlike other values, such as the boundary value \( \delta = 1 \). While this suggests that a single functional observation is sufficient to discriminate between the two classes, unsurprisingly, using all the functions together yields even stronger information.

![Fig 4: Comparison of observation level information for distinguishing between 8-ball and 8-annulus.](image)
5.2. Classification. Here, we assume our data is of the form \((X_i, Y_i) \in \mathbb{R}^d \times \{0, 1\}\). This setting provides the original motivation for considering midpoints \(m_{ij}\) as our anchor points, as in the case of modal clustering, midpoints for within class comparisons would tend to live inside of a cluster (high density, low depth) while between class comparisons would result in midpoints between the clusters (low density, high depth). As we further demonstrate, the way that the depth quantile function transitions between these two extremes provides additional useful information for classification, even outside of the case of convex clusters (cf. section 5.2.1).

To perform binary classification, we consider the following heuristic. We summarize each observation via two functions:

\[
\hat{q}_i^k(\delta) = \frac{1}{\# \{y_j = k\} - \mathbf{1}(y_i = k)} \sum_{j \neq i, y_j = k} \hat{q}_{ij}(\delta), \quad k = 0, 1.
\]

Functional principal component analysis (fPCA) is then used to associate each of these functions with a four dimensional loading vector. Doing this for all functions associates an 8-dimensional vector with each observation, with the first 4 values corresponding to \(\hat{q}_i^0(\delta)\). An ‘out-of-the-box’ support vector machine is used to perform classification on these vectors, with results given below. All other methods presented below for comparison were optimized via leave-one-out cross-validation. Extensions to \(m\)-class problems can be accomplished similarly by constructing \(4m\)-vectors associated with every observation.

5.2.1. Simulated Data. The first classification data set we consider suggests that convex modal regions are not necessary for successfully extracting information relevant for classification with our method. Figure 5 considers a 2-dimensional data set in which the supports of the two classes are a disc and a concentric annulus, such that a third annulus exists between the two with density zero.

Comparisons between classes tend to yield midpoints in this zero density region, and thus valuable information seems to be revealed by small \(\delta\) values. The mapping \(\phi((x_1, x_2)) = (x_1, x_2, x_1^2 + x_2^2)\) makes the data linearly separable though causes the support to live on an elliptic paraboloid and thus all midpoints live in a zero density region. Interestingly, it is the intermediate values of \(\delta\) that now reveal useful information (see figure 5c). Leave-one-out classification resulted in perfect assignments in both the data space and the feature space.

5.2.2. Real Data. We next illustrate the performance of our classification routine on several well known data sets. For each data set, all variables are
first $z$-scaled (all dimensions have mean 0 and variance 1). First is Fisher’s Iris data, consisting of 3 classes of $n_i = 50$ observations each, with $x_i \in \mathbb{R}^4$ (note that we actually perform the classification task in double the dimension in this example). We evaluate the leave-one-out performance of the method, and find a correct classification rate of 97.33%. For comparison, a $k$-nearest neighbors ($k$NN) classifier ($k = 13$) achieves a rate of 96.67% on the $z$-transformed data. Figure 6a demonstrates that not only are the two points clouds situated apart from each other, but they seem to have fairly different geometry as well.

The second real data example we consider is the Wine data set, available from the UCI Machine Learning database. There are 178 observations over 3 classes in 13 dimensions. Our method yields a leave-one-out correct classification rate of 96%. A $k$NN ($k=25$) achieves a rate of 97.7%, though other classifiers have shown even better performance, see Aeberhard et al. (1992). Figures 6b and 6c show the functions for the easiest pairwise comparison (1 misclassification) and the hardest (3 misclassifications), respectively.

We next consider a gene expression data set used for colon cancer detection (Alon et al., 1999), consisting of 62 observations on 2000 variables. We use an opening angle $\alpha = 5\pi/9$ for our cones. Plots of the depth quantile functions (figure 7) seem to suggest that the observations corresponding to the presence of cancer live in a higher density region (early initial dominance of the purple curve) near the boundary of the point cloud (small depth values, despite nearly 2/3 of the observations corresponding to this class). Averages over classes are added to better show the information contained by the collection of functions. A visualization of the first two loading vectors of an fPCA yields more structure for the classification problem than a standard
PCA applied to the raw data. The method returns a correct classification rate of nearly 84% (82% if angle of $\pi/2$ is used). $k$NN ($k = 13$) correctly classifies only 74.6%. A better comparison is to a linear SVM (88.7% with $C = .002$), which demonstrated the best performance in the simulation study of Dutta et al. (2016) under a different evaluation scheme.

Finally, as an infinite dimensional example, we study Spellman’s yeast cell expression data (Spellman et al., 1998), a continuous valued time series. We subsample the data so that only the first $n_t=20$ observations are available for each of 5 classes. Figure 8 shows the raw data, as well as a PCA plot of the data. The second panel considers depth quantile functions for two comparisons, one suggested by PCA to be easy (G1 vs M), and a second suggested to be difficult (S vs G2). The G1 vs M comparison yields a correct
(a) First two components of PCA (left) and raw data (right) for Spellman yeast cell data. Bold curves in right plot are group averages. (right)

(b) Averaged depth quantile functions of easy (G1 vs M) and hard (S vs G2) classification problems - blue, purple are within class comparisons, thicker lines are group averages.

Fig 8: From PCA, G1 and M (red vs cyan in PCA plot) are easy to discriminate between, while S and G2 (green vs blue) are more difficult.

classification rate of 97.5%, while S vs G2 is only 75%. For comparison, we consider the task of discriminating between G1 and non-G1 phases, as explored in Leng and Müller (2006). We find a classification rate of 94% compared to theirs of 90%, though not directly comparable due to different subsetting of the available data.

The method detailed here contains many choices and associated tuning parameters beyond the construction of the depth quantile functions themselves: in particular the number of loadings retained from the fPCA, and the parameters associated with the SVM method (defaults were used above) as well as the averaging of the individual depth quantile functions themselves. How best to use the information is left as future work. However, it is clear that interesting information is available, as the results of this ad hoc method tend to be generally competitive with existing methods, while additionally allowing a visual representation of the data. This visual aspect is particularly beneficial in the next task.

5.3. Anomaly Detection. We conjecture that anchor points associated with outlying observations will tend to live in regions that have different geometry than inlying points. We consider two data sets to illustrate how depth quantile functions are useful in identifying these anomalous points, strongly aided by the visual nature of this representation of the data.

The first data set, the “multiple features” data set (mfeat) considered in Pham (2018) and available on the UCI machine learning repository, consists...
of \( n = 440 \) observations living in \( d = 649 \) dimensions. Each dimension constitutes a different feature of handwritten numerals. This data set consists of 200 patterns each of the numbers ‘6’ and ‘9’ and 40 randomly selected ‘0’ patterns. To illustrate the effect of cone angle, we present the depth quantile functions using both an angle of \( 5\pi/6 \) and \( \pi/2 \) (figure 9). The similarity of the information between the two angles suggest that this is not necessarily a crucial ingredient. Note that the location of the increase in the functions happens at different quantiles depending on angle (\( G \) is the same for both).

Using only a single value of these functions with \( \alpha = 5\pi/6 \) on the normalized functions (so only looking at shape), specifically \( \hat{q}_i(.17)/\hat{q}_i(1) \), yields an area under the response operator curve (ROC AUC) of 0.98, besting the results of Pham (2018) of 0.95, though not directly comparable due to the random subsetting of the data. Another depth-based outlier detection procedure, the kernelized spatial depth of Chen, et al. (2008) only managed 0.76 using a RBF kernel (\( \sigma = 23 \)).

Next, we consider the Lymphography data set, also available at the UCI repository. The data set consists of 4 classes over 148 observations in 18 dimensions. Classes 1 and 4 only constitute 6 instances and are treated as outliers.

Figure 10 shows the analysis of this data using an angle of \( \pi/2 \). The depth quantile functions suggest there are 19 observations (all 6 outliers and 13 inlying points) that warrant further inspection. On inspection of the normalized functions, the outlying functions distinguish themselves from all but one of the inlying functions. Interestingly, the variables in this data set are all numeric assignments of categorical variables, though clearly exploration of the geometry of the induced point cloud is still valuable. This seems to be consistent with Campos et al. (2016), who found an ROC AUC of 1 is
Fig 10: $q_i^k(\delta)$ for lymphography data (left) and their normalized versions of the 19 ‘interesting’ observations (right). Bold solid lines represent the 6 total observations from classes 1 and 4 (black and blue respectively). Dashed lines are for inlying points, color coded for class.

attainable. Chen et al.’s (2008) method returns a value of 0.97 ($\sigma = 16$).

6. Other conceptual relations.

6.1. Interpreting average depth quantiles via random set theory. In an attempt to obtain further insight about the average depth quantiles, we use random set theory to provide a different point of view to the construction of our feature functions. For simplicity we fix $(ij) = (12)$.

Recall that $A_{12}(s) = A_{m_{12},u_{12}}(s)$, $B_{12}(s) = B_{m_{12},u_{12}}(s)$ with $m_{12} = \frac{X_1 + X_2}{2}$ and $u_{12} = \frac{X_1 - X_2}{\|X_1 - X_2\|}$. Let $d_{12}(s) = \min \left( F(A_{12}(s)), F(B_{12}(s)) \right) = F(C^*_{12}(s))$, with $C^*_{12}(s) \in \{A_{12}(s), B_{12}(s)\}$. The joint distribution of $X_1, X_2$ induces the distribution of the random set $C^*_{12}(s)$, and thus of $d_{12}(s)$, the probability content of the set. We can write

$$d_{12}(s) = P_Z(Z \in C^*_{12}(s)|X_1, X_2),$$

where $Z \sim F$ independent of $X_1, X_2$. The corresponding random depth quantile function is $q_{12}(\delta) = d_{12}(s_{12}(\delta))$, where $s_{12}(\delta) = s^*_{12}(\delta)$ is as in Lemma 2.2 but with $(x, u)$ replaced by $(m_{12}, u_{12})$ (equivalently, we could use $s^*_1(\delta)$ instead of $s^*_{12}(\delta)$). Mimicking the construction of the average empirical feature functions used in the classification context (see section 5.2), we obtain by conditioning on $X_1$ and taking expectation over $X_2$ that

$$E_{X_2}[q_{1,2}(\delta)|X_1 = x_1] = E_{X_2}P_Z(Z \in C^*_{12}(s_{12}(\delta))|X_1 = x_1) = E_Z\psi_{x_1,\delta}(Z),$$

18
with \( \Psi_{x_1, \delta}(z) = P_{X_2}(z \in C_{12}^*(s_{12}(\delta))|X_1 = x_1) \) the hitting function of the random set \( C_{12}^*(s_{12}(\delta)) \) conditional on \( X_1 = x_1 \). This random function in \( \delta \) depends on the distributions of both \( X_1 \) and \( X_2 \). In the context of binary classification (with a symmetric \( G \)), this gives three cases corresponding to three different comparison functions \( \delta \rightarrow E_Z \Psi_{X_1, \delta}(Z) \): one between and two within class comparisons. Note that our empirical average depth quantile functions \( \hat{q}_{12}^k(\delta) \) with \( y_j, y_k \in \{0, 1\} \) (see section 5.2) can be considered as samples from the (estimated) distributions of \( E_Z \Psi_{X_1, \delta}(Z) \) for different combinations of class labels of \( X_1, X_2 \). In order for our classification procedure to work, these distributions of the random functions need to differ. Notice that a similar interpretation holds for \( \hat{q}_{12}(\delta) \) by replacing \( P_Z \) by the empirical distribution of the sample \( X_3, X_4, \ldots, X_n \), and the distribution of \( X_2 \) by the empirical distribution of the sample.

6.2. Depth and hitting functions in the literature. Hitting functions have appeared in the context of data depth before, but in a different fashion. For instance, type A depth functions are defined in Zuo and Serfling (2000) as being of the form \( d(x, F) = Eh(x; X_1, \ldots, X_r) \) for some function \( h \), where \( X_i \sim F, i = 1, \ldots, r \). When \( h(x; X_1, \ldots, X_r) = 1(x \in S(X_1, \ldots, X_r)) \), for some (random) set \( S = S(X_1, \ldots, X_r) \), we obtain the depth function as a hitting function \( P(x \in S(X_1, \ldots, X_r)) \). The perhaps best known such depth measure is the simplicial depth introduced by Liu (1988, 1990), where with \( X_i \in \mathbb{R}^d \) and \( r = d + 1 \), \( S(X_1, \ldots, X_{d+1}) = \text{conv}[X_1, \ldots, X_{d+1}] \). Here \( \text{conv}[x_1, \ldots, x_{d+1}] \) is denoting the convex hull of the points \( x_k, k = 1, \ldots, d+1 \). Another instance is the depth based on pairwise connecting spheres (see Elmore et al., 2006) for data on a sphere, and recently extended to data on Riemannian manifolds in Fraiman et al. (2019). Here, given a Riemannian manifold \( \mathcal{M} \) with geodesic distance \( d_g \), a point \( x \in \mathcal{M} \) is assigned the depth \( d(x) = P(x \in B_{X_1, X_2}) \), where \( X_1, X_2 \) are iid, and \( B_{x_1, x_2} \) denotes the ball with radius \( d_g(x_1, x_2) \), centered at the midpoint between \( x_1 \) and \( x_2 \). The work of Ting et al. (2013) is based on a similar premise. Considering data in \( \mathbb{R}^d \), the authors propose to form binary partitions by using random, axis parallel splits. For a given point \( x \in \mathbb{R}^d \), the average mass of the rectangles containing \( x \), taken over a number of binary partitions is considered. It also defines some type of depth function, but the average of the probability contents over a random set is considered, rather than counting the number of boxes containing \( x \). As the discussion above shows, in our work, expected hitting functions come into play when averaging depth quantile functions, and we in fact consider expected hitting functions for an entire class random sets, parametrized by \( \delta \).
6.3. Relations to local depth. Notions of local depth by construction give rise to ‘multiscale’ approaches (e.g. Serfling, 2019). Given any such notion of local depth corresponding a scale parameter \( \tau > 0 \), and given a point \( x \), one can consider a depth function \( \tau \rightarrow d(x, \tau) \). If the scale parameter were chosen according to a distribution \( G \), then this would also allow construction of a depth quantile function corresponding to these local depths similar to what is discussed above. However, one would obtain only one depth quantile function for a given \( x \). In our set-up, however, the notion of depth also depends on the direction, leading to more refined information. For instance, if one were to choose a data point \( X_i \) as an anchor point (cf. see discussion on choice of base point given below), then our construction would result in \( n - 1 \) quantile functions, each describing a different distribution of depths for \( X_i \). It might be interesting to study differences in the distributions of these \( n - 1 \) functions over varying \( X_i \)'s. Also, not all the existing notions of local depth lead to localized information (density) for small scales, as, for instance, the multivariate local simplicial depth, or the local halfspace depth in Agostinelli and Romanazzi (2011) and Agostinelli (2018), respectively.

7. Discussion of choice of tuning parameters. The proposed procedure has three ingredients that can be considered as tuning parameters. The most obvious one is the angle \( \alpha \) of the cones, then there is the distribution \( G \) under which to choose the distance of the tip of the cone to the base point, and last, but not least, there is the choice of the base point. We briefly discuss some aspects of the choice of these ingredients. Moreover, one might also consider our choice of (spherical) cones in the construction of our feature functions to be a ‘tuning parameter’, and we give some discussion on possible alternative choices below.

**Choice of base point:** In this paper, we only consider anchor points of the form \( m_{ij} = \frac{X_i + X_j}{2} \) in defining \( \tilde{q}_{ij}(s) \). It may seem surprising that we choose to describe an observation not by the geometry associated with the observation itself but rather with its associated midpoints. A motivation of this choice in the classification case has been given in the application section, though we use it for all statistical tasks. The rational is that using our depth-based approach with high dimensional point clouds necessitates consideration of values other than the actual observations. It is well known (see Rényi and Sulanke, 1963) that in high dimension, many points will live on or near the convex hull. For such points \( x \), one of the two sets \( A_{x,u}(s) \) or \( B_{x,u}(s) \) will be empty for many if not all values of \( s \), leading to degenerate, non-informative depth quantile functions. Empirically, we have found this to be the case, with many estimated depth quantile functions being identically
zero when considering observations as the anchor points. Considering anchor points that live inside the convex hull, which midpoints often are, helps in alleviating this problem. While other choices might be reasonable as well, we have not explored them here.

Choice of opening angle $\alpha$. As a general guideline, a reasonable choice of $\alpha$ will be such that not all the feature functions will increase very steeply at about the same value of $\delta$, because this will result in weak discrimination. For $G$ fixed, a too large value of $\alpha$ will result in a steep increase early (small $\delta$), while for too small values of $\alpha$ this will happen late. In practice we observed that the shape of the feature functions is not too sensitive to the choice of $\alpha$, adjusting the support of $G$ accordingly. In our simulation studies the choice of an angle of $\pi/2$ seems to work well for a large range of dimensions.

Still, one would expect that the opening angle of the cones should depend on the dimension $d$. In fact, as $d$ increases, the opening angle needs to tend to $\pi$ to avoid the angle being ‘too small’. This is due to the well-known fact that for a fixed opening angle (and tip in 0, say), the portion of the Euclidean space contained in the cone will tend to zero as the dimension increases to infinity. For non-Euclidean data the choice of $\alpha$ appears more challenging, because of some lack of geometric intuition.

As suggested by one of the referees, the angle could also be chosen randomly, similar to the tip $s$. This would mean that one has to choose a two dimensional distribution on $(s, \alpha)$. The resulting depth quantile function would still be a one-dimensional function (because the depth is a one-dimensional quantity), but computationally this would be more challenging, for the depth functions would now be a function of two variables, meaning that its level sets are no longer (one-dimensional) intervals. This will complicate the computation. We will explore this further in future research.

Choice of the distribution $G$. Choosing $G$ essentially means choosing a non-linear transformation to the horizontal axis, amounting to emphasizing different regions of the depth quantile plots. This is somewhat convoluted to the choice of the opening angle $\alpha$ (see above). Choosing $G$ to be unimodal about 0, for instance, would put higher weight on points close to the basepoints, and would thus lead to depth quantile functions that are flatter for small delta than for a uniform $G$, for instance. We recommend to choose $G$ to be symmetric about zero because of the symmetry property of the quantile functions given in Lemma 2.2 part (b). In our applications, we always chose $G$ to be the uniform distribution on an interval chosen such that all the projections used to determine the empirical depth quantiles fall into the support of $G$. The motivation for this choice is similar to the choice of a non-informative prior.
Choice of geometric shapes. Using spherical cones in our procedure is a choice of convenience. Choosing elliptical cones (with the ellipse perhaps chosen adaptively) appears to be a first option. However, choosing an ellipse that works well for all directions $u$ does not appear to be straight-forward. Our current choice of spherically symmetric cones does imply no preference for a direction. Geometric objects with different shapes, such as parabolas or hyperbolas, can also be used. See for instance Kotic and Hlubinka (2017) for a discussion on using such objects in the context of weighted depth. These choices would be similar to cones in that they also provide some sense of direction. In how much the choice of the shapes will effect the performance of the methodology needs to be explored.

8. Properties of depth functions and depth quantile functions. Here we study various properties of the appropriately normalized quantities $\tilde{d}_{x,u}(s)$, $\tilde{q}_{x,u}(\delta)$, $\tilde{d}_{ij}(s)$, and the averaged feature functions $\frac{1}{n} \sum_{j=1}^{n} \tilde{q}_{ij}(\delta)$ considered as processes in $s$ and $\delta$, respectively. On the one hand, the presented results are meant to highlight some interesting properties of these quantities related to the curse of dimensionality and adaptivity, and on the other hand they provide the basis for the formal construction of inference procedures. We do not present such constructions here. The main emphasis of this work is to introduce the basic approach and to illustrate its versatility.

In the following, the opening angle $\alpha$ and the dimension $d$ remain fixed constants unless a dependence on $n$ is indicated explicitly.

8.1. Concentration and large sample behavior. Recall that $\ell_{x,u} = \{ y \in \mathbb{R}^d : y = x + su, s \in \mathbb{R} \}$. For $C, D \subset \mathbb{R}^d$ measurable, define the pseudometric $d_F(C, D) = F(C \Delta D)$, where $C \Delta D = (C \setminus D) \cup (D \setminus C)$. We let $\mathcal{C}_{x_0,u_0} = \{ C_{x,u_0}(s), x \in \ell_{x_0,u_0}, s \in \mathbb{R} \}$ be the set of all cones (with opening angle $\alpha$) with $\ell_{x_0,u_0}$ as their axis of symmetry. With $\mathcal{H}$ the class of all halfspaces with normal vector $u$, $\mathcal{D}_{x_0,u_0} := \mathcal{C}_{x_0,u_0} \cap \mathcal{H}_{u_0}$ equals the set $\{ A_{x,u_0}(s), B_{x,u_0}(s), x \in \ell_{x_0,u_0}, s \in \mathbb{R} \}$. Both $\mathcal{C}_{x_0,u_0}$ and $\mathcal{H}_{u_0}$ are VC-classes with VC-dimension 2 (both are unions of two nested classes of sets), and thus $\mathcal{D}_{x_0,u_0}$ is a VC-class with VC-dimension bounded by 4. The importance here is that the VC-dimension of $\mathcal{D}_{x_0,u_0}$ does not vary with the dimension, which is the reason for the estimators of the depth quantile functions not suffering from the curse of dimensionality, as is shown in the following result. Furthermore, if we let $\mathcal{D} = \bigcup_{(x,u) \in \mathbb{R} \times S^{d-1}} \mathcal{D}_{x,u}$,
then, with $C = Sx^2R$, $u^2Sd^1Cx,u$ and $H = Su^2Sd^1Hu$, we have $D = \{ C \cap H, C \in C, H \in H \}$. $D$ is a VC-class, since both $C$ and $H$ are VC-classes. However its VC-dimension depends on the dimension $d$. In fact, it is $O(d)$.

8.1.1. Results for empirical depth functions.

**Proposition 8.1** For every given line $\ell_{x_0,u_0}$, there exists a constant $c > 0$ not depending on the dimension $d$ and the opening angle $\alpha$ such that, for all $M > 0$,

$$P\left[ \sup_{x \in \ell_{x_0,u_0}, s \in \mathbb{R}} \left| \sqrt{n}(\hat{d}_{x,u_0}(s) - d_{x,u_0}(s)) \right| > M \right] \leq cM^5 \exp\left\{ -2M^2 \right\}. \quad (8.1)$$

Now recall that our methodology depends on the $b_{dij}(s)$ functions, $i < j$. We show that, with $A_{ij}(s) = A_{m_{ij},u_{ij}}(s)$ and $B_{ij}(s) = B_{m_{ij},u_{ij}}(s)$, these functions are uniformly close to the following population based functions

$$d_{ij}(s) = \min(F(A_{ij}(s)), F(B_{ij}(s))).$$

While $d_{ij}(s)$ depends on the population distribution, it is still a random quantity, but only through the choice of the line and the base point. We now have the following result:

**Proposition 8.2** With $d = d_n \in \mathbb{N}$ and $\alpha = \alpha_n \in [0, \pi)$, both nonrandom sequences, we have,

$$\max_{i,j=1,\ldots,n} \sup_{s \in \mathbb{R}} |\hat{d}_{ij}(s) - d_{ij}(s)| = O_P\left( \sqrt{\frac{\min(d, \log n)}{n}} \right) \quad \text{as } n \to \infty. \quad (8.2)$$

Moreover, if the distribution $F = F^{(n)}$ (as $d$ might depend on $n$, the same holds for $F$) is concentrated on a $d'$-dimensional hyperplane in $\mathbb{R}^d$, then (8.2) holds with $d$ replaced by $d'$, where $d'$ might also depend on $n$.

**Remarks.** (i) In case $d = d_n \to \infty$, in order for the proposition to give non-trivial results, the angle $\alpha = \alpha_n$ needs to tend to $\pi$ at a specific rate depending on $d$ (assuming the density $f = f^{(n)}$ of $F = F^{(n)}$ stays bounded), for otherwise the area of the $A$-sets and the $B$-sets, and thus also $\hat{d}_{ij}(s)$ and $d_{ij}(s)$, will tend to zero. To work out explicit conditions on the behavior of $F^{(n)}$, $d_n$ and $\alpha_n$ that avoid degeneracies is technically involved. (ii) The ‘min’ appearing in the rate tells us that $\left( \frac{\log n}{n} \right)^{1/2}$ is a dimension independent uniform bound for the rate of the deviations of the functions $\hat{d}_{ij}(s)$ from their population based counterparts $d_{ij}(s)$. The asserted adaptivity to some kind of ‘sparsity’ is due to some geometry: Consider a cone in $\mathbb{R}^d$ with
tip at zero and with axis of symmetry being one of the coordinate axes. The intersection of such a cone with a linear subspace of dimension $d'$ is a cone in $\mathbb{R}^{d'}$. In case $d = d_n \to \infty$, in order for the proposition to give non-trivial results, the angle $\alpha = \alpha_n$ needs to tend to $\pi$ at a specific rate depending on $d$ (assuming the density $f = f^{(n)}$ of $F = F^{(n)}$ stays bounded), for otherwise the area of the $A$-sets and the $B$-sets, and thus also $d_{ij}(s)$ and $d_{ij}(s)$, will tend to zero. To work out explicit conditions on the behavior of $F^{(n)}$, $d_n$ and $\alpha_n$ that avoid degeneracies is technically involved.

Our next result concerns the asymptotic distribution of $\sqrt{n}(\hat{d}_{x,u}(s) - d_{x,u}(s))$ as a process in $s \in \mathbb{R}$. The pointwise behavior of the process depends on whether the minimizing set in the definition of $d_{x,u}(s)$ is unique or not (see Corollary 8.1). Therefore, we are facing a challenge in small neighborhoods of points with non-unique minimizers, because the empirical minimizer, i.e. the sets minimizing $\hat{d}_{x,u}(s)$, might not stabilize in such neighborhoods. Moreover, suppose that $s_0$ is a point with $d_{x,u}(s_0) = F(A_{x,u}(s_0)) = F(B_{x,u}(s_0))$, and consider $s$ in a small neighborhood of $s_0$. If the minimizers corresponding to $s < s_0$ are $A_{x,u}(s)$, and for $s' > s_0$ they switch to $B_{x,u}(s')$, then the map $s \mapsto 1_{C_{x,u}^s(s)}$ is not continuous at $s_0$ (w.r.t. the $L_1(F)$-norm), because the sets $A_{x,u}(s)$ and $B_{x,u}(s')$ are not nested.

The following definition is to exclude such values $s_0$. For $(x, u) \in \mathbb{R}^d \times S^{d-1}$, denote $\Delta_{x,u}(s) = |F(A_{x,u}(s)) - F(B_{x,u}(s))|$, and for $\eta > 0$ let

$$
(8.3) \quad S_{x,u}(\eta) = \{s \in \mathbb{R} : \Delta_{x,u}(s) \geq \eta\}.
$$

Note that $\Delta_{x,u}(s) = 0$ means that the minimizer $C_{x,u}^s(s)$ in the definition of $d_{x,u}(s)$ is not unique (both $A_{x,u}(s)$ and $B_{x,u}(s)$ are minimizers). We also introduce the notation

$$
T_{x,u}(s) = \arg\min_{C \in \{A_{x,u}(s), B_{x,u}(s)\}} F(C).
$$

Observe that this set of minimizers $T_{x,u}(s)$ either consists of exactly one, or of exactly two elements. The latter holds if and only if $\Delta_{x,u}(s) = 0$.

**Theorem 8.1** For $(x, u) \in \mathbb{R} \times S^{d-1}$ set $S_{x,u}^*(\eta) = S_{x,u}(\eta) \cup \{s : \Delta_{x,u}(s) = 0\}$, and let $\{\eta_n\}$ be a sequence of real numbers with $\eta_n \sqrt{n} \to \infty$. Then, on an appropriate probability space, there exists a sequence of $F$-bridges $\{B_F^{(n)}(C), C \in \mathcal{D}\}$, i.e. tight, mean zero Gaussian processes with covariance $\text{Cov}(B_F^{(n)}(C), B_F^{(n)}(D)) = F(C \cap D) - F(C)F(D)$, and almost surely continuous sample paths w.r.t. $d_F$, such that, for any given line $\ell_{x,u}$, we have with
\[ B_{x,u}(s) = \min \{ B_{F}^{(n)}(C), C \in \mathcal{T}_{x,u}(s) \} \] that
\[
\sup_{s \in S_{x,u}^*} \left| \sqrt{n}(\widehat{d}_{x,u}(s) - d_{x,u}(s)) - B_{x,u}^{(n)}(s) \right| = o_P(1), \quad \text{as } n \to \infty,
\]
where the \( o_P(1) \)-term does not depend on \( d, \alpha \) and \( u \). Moreover, we have
\[
\sup_{(x,u) \in \mathbb{R} \times S_{x,u}^*} \left| \sqrt{n}(\widehat{d}_{x,u}(s) - d_{x,u}(s)) - B_{x,u}^{(n)}(s) \right| = o_P(1), \quad \text{as } n \to \infty,
\]
where the \( o_P(1) \)-term does not depend on \( \alpha \), but it does depend on \( d \).

**Remark.** In the somewhat related context of minimum volume sets (or generalized quantile processes), a result of similar type has been obtained in Einmahl and Mason (1992). There the approximating process (to the generalized quantile process) is a maximum of \( F \)-bridges taken over the, in general non-unique, generalized quantiles. Also, Massé (2004) showed convergence of the multivariate Tukey depth process to a similar type of limit process.

**Corollary 8.1** For \((x,u) \in \mathbb{R}^d \times S^{d-1}\), set \( \sigma_{x,u}^2(s) = d_{x,u}(s)(1 - d_{x,u}(s)) \), \( s \in \mathbb{R} \).

(i) For \((x,u,s) \in \mathbb{R}^d \times S^{d-1} \times \mathbb{R} \) with \( \Delta_{x,u}(s) \neq 0 \), we have
\[
\sqrt{n}(\widehat{d}_{x,u}(s) - d_{x,u}(s)) \to \mathcal{N}(0, \sigma_{x,u}^2(s)) \quad \text{in distribution as } n \to \infty.
\]

(ii) For \((x,u,s) \in \mathbb{R}^d \times S^{d-1} \times \mathbb{R} \) with \( \Delta_{x,u}(s) = 0 \), we have
\[
\sqrt{n}(\widehat{d}_{x,u}(s) - d_{x,u}(s)) \to \min(N_1, N_2) \quad \text{in distribution as } n \to \infty,
\]
where \( N_1, N_2 \) are jointly normal with mean zero, variance \( \sigma_{x,u}^2(s) \), and covariance \( \text{Cov}(N_1, N_2) = -d_{x,u}^2(s) \).

8.1.2. Results for empirical depth quantile functions. First we give some finite sample concentration results.

**Proposition 8.3** For every fixed line \( \ell_{x_0,u_0} \subset \mathbb{R}^d \), there exists a constant \( c > 0 \), not depending on \( d \) and \( \alpha \), such that, for all \( M > 0 \),
\[
P\left[ \sqrt{n} \sup_{x \in \ell_{x_0,u_0}, \delta \in [0,1]} \left| \widehat{q}_{x,u_0}(\delta) - q_{x,u_0}(\delta) \right| > M \right] \leq cM^3e^{-M^2/2}.
\]

Our next result addresses the asymptotic distribution of the depth quantile functions. We need the following assumptions.

(A1) The distributions \( F \) and \( G \) have bounded, continuously differentiable
Lebesgue densities \( f > 0 \) on \( \mathbb{R}^d \), and \( g > 0 \) on \( \mathbb{R} \), respectively.

\textbf{(A2) (tail behavior of \( f \))} There exist constants \( C, \epsilon > 0 \) such that

\begin{align*}
(\text{i}) \quad & \sup_{z \in \mathbb{R}^d} \|z\|^{d-1+\epsilon} f(z) < C < \infty \quad \text{and} \\
(\text{ii}) \quad & \sup_{z \in \mathbb{R}^d} \|z\|^{d-1+\epsilon} \|\nabla f(z)\| < C < \infty.
\end{align*}

\textbf{(A3) For each} \( K > 0 \), there \textit{exists} \( \epsilon_K > 0 \) \textit{such that}

\[ \inf_{\|x\| \leq K, u \in S^{d-1}} \inf_{s, \Delta_{x,u}(s) = 0} \left| \frac{d}{ds} F(A_{x,u}(s)) - \frac{d}{ds} F(B_{x,u}(s)) \right| > \epsilon_K, \]

where \textbf{(A1)} and \textbf{(A2)} imply \textit{these} \textit{derivatives} \textit{exist} \textit{(see Lemma 0.2 in the supplemental material).}

\textbf{Discussion of (A3).} This assumption is used to show that, for each \((x,u)\) and \( \eta \), the set \( S_{x,u}^C(\eta) \) consists of a (finite) union of ‘small’ intervals, and that the same is true for is the set \( D_{x,u}^C(\eta, \epsilon) \); cf. Lemma 0.3. Intuitively, assumption \textbf{(A3)} can be expected to hold in many situations. The reason for this intuition is that \( S_{x,u}^C(\eta) \) can only be large, if \( F(A_{x,u}(s)) \) and \( F(B_{x,u}(s)) \) have the same derivatives (w.r.t. \( s \)) in a neighborhood of a point \( s_0 \) with \( F(A_{x,u}(s_0)) = F(B_{x,u}(s_0)) \). These derivatives are surface integrals (with respect to \( f \)) over the ‘lateral’ surface areas of the cone \( A_{x,u}(s) \) and the frustum \( B_{x,u}(s) \), respectively (see proof of Lemma 0.2). Given the geometry of these sets, these derivatives being equal over a (small) interval, puts some strong requirements on the geometry of the density \( f \). For more standard density models, these requirements do not seem to hold, at least intuitively, while explicit calculations appear challenging, even for a normal model, say. Also, suppose that while \( F(A_{x,u}(s_0)) = F(B_{x,u}(s_0)) \), the derivative of these two functions at \( s_0 \) is different, let’s say, \( \frac{d}{ds} F(A_{x,u}(s_0)) > \frac{d}{ds} F(B_{x,u}(s_0)) \). Then, in order for \( F(A_{x,u}(s_1)) = F(B_{x,u}(s_1)) \) to hold for some \( s_1 > s_0 \), the inequality between the derivatives has to change along the way. Again, this puts some requirements on the geometry of the underlying density. Intuitively, these requirements are more likely to hold, if \( f \) has several modes, allowing for a less regular behavior of these derivatives, by which we mean that they have a large number of local extrema. This leads to the intuition that \( S_{x,u}^C(\eta) \) (and \( D_{x,u}(\delta) \)) can be expected to consist of a finite number of small intervals assuming that the number of modes of \( f \) is finite.

Recall the notation \( \{s : d_{x,u}(s) \leq t\} = [s_{x,u}^l(t), s_{x,u}^r(t)] \), and that for \( t = q_{x,u}(\delta), 0 \leq \delta < 1 \), we use the short-hand notation

\[ \{s : d_{x,u}(s) \leq q_{x,u}(\delta)\} = [s_{x,u}^l(\delta), s_{x,u}^r(\delta)]. \]
Also recall the definition of the set $S_{x,u}(c)$ given in (8.3), and that $C^*_x(s) \in \{A_{x,u}(s), B_{x,u}(s)\}$ denotes the minimizer in the definition of $d_{x,u}(s)$. We now have the following result.

**Theorem 8.2** Suppose that assumptions (A1) - (A3) hold. For $C, \epsilon, \eta > 0$, let $D_{x,u}(\eta, \epsilon) = \{\delta \in [\epsilon, 1 - \epsilon] : s^r_{x,u}(\delta), s^l_{x,u}(\delta) \in S_{x,u}(2\eta)\}$. For each $K > 0$, and for real numbers $\eta_n > 0$ with $\eta_n \to 0$ and $\eta_n \sqrt{n} \to \infty$, we have, as $n \to \infty$,

$$
\sup_{u \in S} \sup_{\|x\| \leq K} \left| \sqrt{n}(\tilde{q}_{x,u}(\delta) - q_{x,u}(\delta)) - \left[ \alpha_{x,u}(\delta) \nu_n(s^l_{x,u}(\delta)) - (1 - \alpha_{x,u}(\delta)) \nu_n(s^r_{x,u}(\delta)) \right] \right| = o_p(1),
$$

where $\alpha_{x,u}(\delta) = (1 + \frac{g(s^r_{x,u}(\delta))}{g(s^l_{x,u}(\delta))} \cdot \frac{d_{x,u}(s^l_{x,u}(\delta))}{d_{x,u}(s^r_{x,u}(\delta))})^{-1}$, and $\nu_n(s_{x,u}(\delta)) = \sqrt{n}(F_n - F)(C^*_x(s_{x,u}(\delta)))$. Moreover, on an appropriate probability space, there exists a sequence of mean zero $F$-bridges $\{\mathbb{E}_F^{(n)}(C_{x,u}(s))\}$, such that with $\mathbb{E}_{x,u}^{(n)}(s) = \mathbb{E}_F^{(n)}(C_{x,u}(s))$, we have, as $n \to \infty$,

$$
\sup_{u \in S} \sup_{\|x\| \leq K} \left| \sqrt{n}(\tilde{q}_{x,u}(\delta) - q_{x,u}(\delta)) - \left[ \alpha_{x,u}(\delta) \mathbb{E}_{x,u}^{(n)}(s^l_{x,u}(\delta)) - (1 - \alpha_{x,u}(\delta)) \mathbb{E}_{x,u}^{(n)}(s^r_{x,u}(\delta)) \right] \right| = o_p(1).
$$

**Discussion:** (i) As $\sqrt{n}(\tilde{q}_{x,u}(\delta) - q_{x,u}(\delta))$ is a quantile-type process, and the approximating process is a (mixture of two) empirical processes given by $\nu_n(s^l_{x,u}(\delta))$ and $\nu_n(s^r_{x,u}(\delta))$, respectively, the first assertion can be seen as a generalized Bahadur-Kiefer representation.

(ii) The approximating Gaussian processes $\mathbb{E}_{x,u}^{(n)}(s_{x,u}(\delta))$ depend on $C^*_x(s)$, which is either $A_{x,u}(s)$ or $B_{x,u}(s)$. In other words, either $\mathbb{E}_{x,u}^{(n)}(s_{x,u}(\delta)) = \mathbb{E}_F^{(n)}(A_{x,u}(s_{x,u}(\delta)))$ or $\mathbb{E}_{x,u}^{(n)}(s_{x,u}(\delta)) = \mathbb{E}_F^{(n)}(B_{x,u}(s_{x,u}(\delta)))$. Both the processes $\{\mathbb{E}_F^{(n)}(A_{x,u}(s_{x,u}(\delta))), \delta \in [0, 1]\}$ are time-transformed one-dimensional Brownian bridges, because both $\{A_{x,u}(s), s \in \mathbb{R}\}$ and $\{B_{x,u}(s), s \in \mathbb{R}\}$ consists of nested classes of sets, and because the functions $\delta \to s^l_{x,u}(\delta)$ are monotonic. As $\delta$ varies, we can think of $\mathbb{E}_{x,u}^{(n)}(s_{x,u}(\delta))$ as switching between these two processes, depending on which of the two sets $A_{x,u}(s)$ or $B_{x,u}(s)$ is the corresponding minimizer of the depth functions. The approximation works for $\delta \in D_{x,u}(\eta_n, \epsilon)$, because we know that the empirical minimizer and the theoretical minimizers coincide with probability tending to one. By our assumption, the set $D_{x,u}(\eta_n, \epsilon)$ is a finite union of intervals, separated
by small neighborhoods about points \( s_0 \) with \( \Delta_{x,u}(s_0) = 0 \), so that there is no switching between approximating processes within a subinterval.

(iii) The mixing proportion \( \alpha_{x,u}(\delta) \) is determined by the two quantities \( g(s_{x,u}^r(\delta))/d_x^r(s_{x,u}(\delta)) \) and \( g(s_{x,u}^l(\delta))/d_x^l(s_{x,u}(\delta)) \), which are the derivatives of \( G(s \geq 0 : d_{x,u}(s) \leq t) \) and \( G(s < 0 : d_{x,u}(s) \leq t) \), respectively, evaluated at \( t = q_{x,u}(\delta) \). The restrictions on \( x \) and \( \delta \) given in the statements of the theorem make sure that these ratios are uniformly bounded away from zero and infinity, and that they are uniformly continuous maps as \( x \) and \( \delta \) vary on the restricted sets. This is needed to assure that the asserted uniformity of the asymptotic statements.

**Corollary 8.2** Under the assumptions of Theorem 8.2, we have for each \( (x,u) \in \mathbb{R}^d \times S^{d-1} \) and \( \delta \in [0,1) \) with \( \Delta(s_{x,u}^r(\delta)) \neq 0 \) and \( \Delta(s_{x,u}^l(\delta)) \neq 0 \), we have, as \( n \to \infty \),

\[
\sqrt{n}(\tilde{q}_{x,u}(\delta) - q_{x,u}(\delta)) \to \alpha_{x,u}(\delta)N_{x,u}^l(\delta) + ((1 - \alpha_{x,u}(\delta))N_{x,u}^r(\delta) \text{ in distribution,}
\]

where \( N_{x,u}^l(\delta), N_{x,u}^r(\delta) \) are mean zero, jointly normal random variables with variances \( F(C_{x,u}^*(s_{x,u}^l(\delta))) [1 - F(C_{x,u}^*(s_{x,u}^l(\delta)))] \) and \( F(C_{x,u}^*(s_{x,u}^r(\delta))) [1 - F(C_{x,u}^*(s_{x,u}^r(\delta)))] \), respectively. Their covariance is given by \( \text{Cov}(N_{x,u}^l(\delta), N_{x,u}^r(\delta)) = F(C_{x,u}^*(s_{x,u}^l(\delta)) \cap C_{x,u}^*(s_{x,u}^r(\delta))) - F(C_{x,u}^*(s_{x,u}^l(\delta))) F(C_{x,u}^*(s_{x,u}^r(\delta))) \).

### 8.1.3. Asymptotic normality of averaged depth quantile functions.

Here we study the asymptotic behavior of the averaged depth quantile functions used in the applications. Recall that \( m_{ij} = \frac{X_i + X_j}{2} \) and \( u_{ij} = \frac{X_i - X_j}{\|X_i - X_j\|} \). For \( K > 0 \), let \( \tilde{q}_{i,K}(\delta) = \frac{1}{n} \sum_{j \neq i}^{n} \tilde{q}_{ij}(\delta)1(||m_{ij}\| \leq K) \), \( i = 1, \ldots, n \), and let the corresponding population-based quantities be \( q_{i,K}(\delta) = \frac{1}{n} \sum_{j \neq i}^{n} q_{ij}(\delta)1(||m_{ij}\| \leq K) \), \( i = 1, \ldots, n \). Consider the process

\[
\sqrt{n}(\tilde{q}_{i,K}(\delta) - E(q_{i,K}(\delta)|X_i)) \in [0,1],
\]

conditional on \( X_i \). We will see that this process can be approximated by a U-process, which will then gives us weak convergence to a Gaussian process. Recall the definition of \( D_{x,u}(\eta, \epsilon) \) given in Theorem 8.2, and let \( D_{ij}(\eta, \epsilon) = D_{m_{ij},u_{ij}}(\eta, \epsilon) \). We need to following additional assumption:

**(A4)** Let \( \Psi_{\delta,u}(x) = F(A_{x,u}(s_{x,u}(\delta))) - F(B_{x,u}(s_{x,u}(\delta))) \). The directional derivatives \( \nabla_u \Psi_{\delta,u}(x) \) at \( x \) in direction \( u \in S^{d-1} \) exist and are equi-Lipschitz continuous for \( u \in S^{d-1}, \delta \in [\epsilon, 1 - \epsilon], \epsilon > 0 \). Moreover, for some \( \epsilon_K > 0 \),

\[
\inf_{x \in B^{d-1}} \inf_{\|u\| \leq K} \delta : \Delta_{x,u}(s_{x,u}(\delta)) = 0 \left| \nabla_u \Psi_{\delta,u}(x) \right| > \epsilon_K.
\]
DISCUSSION OF (A4). Validity of assumption (A4) depends on the size and the shape of $D_x^k(\eta, \epsilon)$ as a function in $(x, u)$, which in turn is another implicit assumption of the shape of the underlying density $f$. Suppose we fix a line $\ell$ (in particular $u$ is fixed), and we consider the directional derivative of $F(A_{x,u}(s))$ and $F(B_{x,u}(s))$ for fixed $s$ w.r.t. $x$ in direction $u$. For $s > 0$, the directional derivative of $F(B_{x,u}(s))$ w.r.t. $x$ in direction $u$ is given by the negative of the entire surface integral over the entire surface of the frustum $B_{x,u}(s)$. The corresponding directional derivative of $F(A_{x,u}(s))$ equals the difference of the surface integral w.r.t. $f$ over the ‘lateral’ surface and the base of $A_{x,u}(s)$. The fact that this derivative is a difference makes it even more difficult to investigate than the derivative in (A3). Moreover, in (A4), the value $s$ also is a function of $(x, u)$, which adds to the complexity. Nevertheless, heuristic arguments similar to the ones given in the discussion of (A3) leads to the intuition that assumption (A4) holds in many cases.

Theorem 8.3 Suppose that assumptions (A1) - (A4) hold. Let $\epsilon > 0$. For $K > 0$, we have conditional on $X_1$, as $n \to \infty$, almost surely,

$$\sqrt{n}\left(q_{1,K}(\delta) - E(q_{12}(\delta)1(\|m_{12}\| \leq K)|X_1)\right) \to G_{X_1}(\delta) \text{ weakly, in } D[\epsilon, 1-\epsilon],$$

where conditional on $X_1$, $\{G_{X_1}(\delta), \delta \in [\epsilon, 1-\epsilon]\}$ is a tight Gaussian process with mean zero and $\text{Cov}[G_{X_1}(\delta), G_{X_1}(\delta')]|X_1] = \text{Cov}[(q_{12}(\delta)1(\|m_{12}\| \leq K), q_{12}(\delta')1(\|m_{12}\| \leq K))|X_1].$

REMARKS. (i) Clearly, conditional on $X_i$, the distribution of $\sqrt{n}\left(q_{i,K}(\delta) - E(q_{i}(\delta)|X_i)\right)$ is the same for every $i = 1, \ldots, n$.

(ii) Up to truncation, the conditional mean used for centering equals the quantity $E(X\left(\Psi_{x_1,F_{x_2},\delta}(X)\right))$ introduced in section 6.1, where $X_1 = x_1$ is the observed value of $X_1$. As discussed in section 6.1, the quality of classification based on depth quantile functions as discussed in section 5.2, depends on the behavior of these quantities as a function in $\delta$.

(iii) Theorem 8.3 concerns an average of processes conditional on $X_1$, centered at their (conditional) means $E(q_{12}(\delta)|X_1)$. If we, as above, center the individual processes on $q_{12}(\delta)$ instead, meaning we are subtracting the conditional expectation given $X_2$ (and given $X_1$), then the resulting average process $\frac{1}{n}\sum_{j=2}^{n}\sqrt{n}\left(q_{1,j}(\delta) - q_{1,j}(\delta)\right)1(\|m_{1j}\| \leq K), \delta \in [\epsilon, 1-\epsilon]$ can be approximated by a degenerate $U$-process, which converges at a rate faster than $1/\sqrt{n}$. This is shown in the proof of Theorem 8.3.

(iv) It is interesting to observe that, in contrast to Theorem 8.2, there is no exceptional set. In fact, the exceptional sets ‘average out’ because we are averaging over many different lines. Here is where assumptions (A3) and (A4) come into play, because they make sure that individual exceptional sets are ‘small’. (See proof of Theorem 8.3 for details.)
Acknowledgements. The authors would like to thank the Associate Editor and three reviewers for their very thorough reading of the original manuscript and for many helpful comments that greatly improved readability and accessibility of the material.

9. References.

Aeberhard, S., Coomans, D. and de Vel, O. (1992): Comparison of Classifiers in High Dimensional Settings, Tech. Rep. no. 92-02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.

Agostinelli, C. (2018): Local half-region depth for functional data. *J. Multivariate Anal.*, 163, 67 - 79.

Agostinelli, C. and Romanazzi, M. (2011): Local depth. *J. Statist. Plan. Inference* 141, 817 - 830.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96(12), 6745-6750.

Arcones, M.A., and Giné, E. (1993): Limit theorems for U-processes. *Ann Probab.*, 21, 1494 - 1542.

Berthet, P. and Mason, D. M. (2006): Revisiting two strong approximation results of Dudley and Philipp, In *IMS Lecture Notes - Monograph Series, High Dimensional Probability*, Giné, E., Koltchinskii, V., Li, W., and Zinn, J. eds., 51, 155 - 172.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16, 77-102.

Bubenik, P. and Kim, P. T. (2007). A statistical approach to persistent homology. *Homology Homotopy Appl.*, 9, 337-362.

Bousquet O., Boucheron S., and Lugosi G. (2004): Introduction to Statistical Learning Theory. In: Bousquet O., von Luxburg U., Rätsch G. (eds) *Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science, vol. 3176*. Springer, Berlin, Heidelberg.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. (2016): On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study. *Data Mining and Knowledge Discovery* 30, 891-927.

Chaudhuri, P. and Marron, J.S. (1999): SiZer for exploration of structure in curves. *J. Amer. Stat. Assoc.* 94, 807-823.

Chen, Y., Dang, X., Peng, H., and Bart, H. L. (2008): Outlier Detection with the Kernelized Spatial Depth Function, *IEEE T. Pattern. Anal.*, 30
Cuturi, M. (2010): Positive definite kernels in machine learning. *Technical report.*

Dutta, S., Sarkar, S., and Ghosh, A.K. (2016): Multi-scale classification using localized spatial depth. *J. Machine Learn. Res.*, **17**, 1 - 30.

Einmahl, J.H.J. and Mason, D. (1992): Generalized quantile processes. *Ann. Stat.* **20**, 1062-1078.

Elmore, R.T., Hettmansperger, T.P. and Xuan, F. (2006): Spherical data depth and a multivariate median, In: *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., **72**, 87-101.

Fraiman, R., Gamboa, F. and Moreno, L. (2019): Connecting pairwise geodesic spheres by depth: DCOPS. *J. Multivariate Anal.*, **169**, 81-94.

Genton, M. (2001): Classes of kernels for machine learning: A statistics perspective. *J. Machine Learn. Res.*, **2**, 299-312.

Kotik, L. and Hlubinka, D. (2017): A weighted localization of halfspace depth and its properties. *J. Multivariate Anal.*, **157**, 53-69.

Leng, X. and Müller, H.G. (2006): Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22**, 68-76.

Liu, R.Y. (1988): On a notion of simplicial depth, *Proceedings of the National Academy of Sciences*, **85**, 1732-1734.

Liu, R.Y. (1990): On a notion of data depth based on random simplices, *Ann. Statist.*, **18**, 405-414.

Massart, P. (1989): Strong approximation for multivariate empirical and related processes, via KMO constructions. *Ann. Probab.* **17**, 266 - 291.

Massé, J-C. (2004): Asymptotics for the Tukey depth process with an application to a multivariate trimmed mean. *Bernoulli* **10**, 397-419.

Minotte, M. and Scott, D. (1993): The mode tree: A tool for visualization of nonparametric density features, *J. Comput. Graph. Statist.* **2**, 51-68.

Paindaveine, D. and van Bever, G. (2012): From depth to local depth: A focus on centrality. *JASA*, **108:503**, 1105 - 1119.

Pham, N. (2018): L1-Depth Revisited : A Robust Angle-based Outlier Factor in High-dimensional Space. *Machine Learning and Knowledge Discovery in Databases*, pp 105-121.

Rényi, A. and Sulanke, R. (1963): Über die konvexe Hülle von n zufällig gerwählten Punkten I. *Z. Wahrsch. Verw. Gebiete*, **2**, 75–84.

Serfling, R. (2019): Depth Functions on General Data Spaces, I. Perspectives, with Consideration of “Density” and “Local” Depths. *Available at www.utdallas.edu/~serfling/papers/I.Perspectives.pdf*
Shawe-Taylor, J. and Cristianini, N. (2004): *Kernel methods for pattern analysis*. Cambridge University Press.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998): Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* 9, 3273 - 3297.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000): A global geometric framework for nonlinear dimensionality reduction, *Science*, 290, 2319 - 2323.

Ting, K.M., Zhou, G-T., Liu, F.T., and Tan, S.C. (2013): Mass estimation. *Mach. Learn.*, 90, 127-160.

van der Vaart, A. W. and Wellner, J. A. (1996): *Weak convergence of empirical processes: With applications to Statistics*. Springer

Zuo, Y. and Serfling, R. (2000): Notions of statistical depth functions. *Ann. Statist.*, 28, 461-482.